Region-Specific Modeling of Brain Response to Visual Stimuli

Jeremy Tsai, Department of Computer Science, UCLA

1 Introduction

The Natural Scenes Dataset (NSD Dataset) is a dataset of fMRI readings of 8 subjects as they are presented various natural images. The researchers at Center of Magnetic Resonance Research (CMRR) at the University of Minnesota who oversee this dataset originally planned to present a total of 10000 distinct images three times in a predetermined shuffled order to 8 experiment subjects over the 40 fMRI sessions for each person. Out of the 10000 distinct images, 1000 images were shared across all subjects. However, due to experiment constraints, some of the subjects were not able to take part in all 40 sessions, so for them some images were never shown or was shown less than three times. In particular, only 907 out of the shared set of 1000 images were shown to every subject at least one.[[1]](#endnote-1)

The NeuroGen framework, developed by Zijin Gu with collaborators from Cornell University and University of Minnesota, connects an encoding model and a generative adversarial network (GAN) to synthesize images that can maximize certain areas of the human visual system. In this framework, the GAN and the encoding interacts in a back-and-forth manner. First, the GAN generates a synthetic image from a noise vector. The encoding model then predicts the brain response to the synthetic image. Since the gradient flows “from the encoding model’s predicted response back to the synthetic image and to the noise vector that initializes the conditional GAN”, it is possible to fine tune the noise vector that created the synthetic image to begin with in a way that the new synthetic image would elicit a stronger response in the brain region. The process then repeats itself for a user-specified number of times, after which the best performing image can be obtained.[[2]](#endnote-2)

The encoding model that is a crucial part in the NeuroGen framework is trained using the NSD dataset. The encoding model used in NeuroGen is a fwRF model that can achieve state-of-the-art brain activation prediction accuracy, but it is trained (and only suitable) for predicting only one NSD Dataset subject’s brain response at a time. This subject-specificity is problematic, as the synthetic images created using this subject-specific encoding model would likely be incapable of achieving the theoretical level of brain activation when the synthetic images are shown to subjects other than the one that the encoding model was trained for. As such, an encoding model that can predict the brain activation of a “general” brain would be better for practical use. The goal for this summer project is to explore different ways of modeling the response of an arbitrary brain to an arbitrary image. The underlying assumption of this project is that the average brain activation of the 8 subjects from the NSD Dataset is representative enough of the response of a “general” brain. Although the task of predicting the average brain activation of the 8 subjects can technically be achieved by averaging the predictions made by the 8 subject-specific fwRF models, this would be terribly inefficient, as running 8 models at the same time would be a significant computational burden.

2 Theory

Neural networks are composed of units called “neurons”. To model actual neurons, these neurons are capable of receiving some numerical input and sending an output to other neurons. The output of these neurons is usually just a weighted output of some non-linear function. As the network learns, these weights adjust themselves in a way that it amplifies informative inputs and reduces redundant ones. These neurons are typically organized into layers, which are a group of neurons that takes input from another neurons. For instance, the neurons from the third layer receives signals from the second layer. The special cases, the first and the last layer of the network, are called the input and output layers, and all other layers are called the hidden layers.

Since this encoding model is meant to predict brain’s response to visual stimuli, it would make sense to start off by using some techniques in computer vision. In recent years, convolutional neural networks (CNN) have been extremely successful in computer vision related tasks, such as object recognition and image classification. As opposed to older, simpler neural networks that have fully connected layers, where all neurons of the same layer receive input from all neurons of the preceding layer, CNNs use convolutional layers. In a convolutional layer, each neuron only receives input from a localized group of neurons from the preceding layer. Because of these localized groups, CNNs can encode spatial information.[[3]](#endnote-3) The product of these convolutional layers is called the feature maps, which typically encodes more abstract information such as edges and shades. Intuitively, the human visual system should react to these abstract visual features rather than actual pixel values. Therefore, the decision to use feature maps to predict brain activations was made.

There has been a significant number of CNNs that have achieved remarkable performance with the tasks they are designed for. AlexNet is one of the earliest CNNs that have been extremely successful in the field of image classification. It incorporates 5 convolutional layers followed by 3 fully connected layers. Before the feature maps go from the fifth convolutional layer to the first fully connected layer, the feature maps need to be flattened, which means that from this point on, there will be no further spatial information that are extracted from the image, as fully connected layers are being used.

It is decided that AlexNet would be continued to be used in this project. AlexNet is readily available online, which means that the bulky convolutional network does not have to be delivered along with this project’s relevant scripts, as any computer running the script with an internet connection can obtain the model. Furthermore, AlexNet is shown to be capable of achieving decent results in NeuroGen (it is used to generate feature maps used by the fwRF encoding model), so it is likely to be useful in the development of an average encoding model.

As per Zijin Gu’s advice, since the average encoding model is ultimately a regression model, there should be some investigation to linear models in this task. This is because linear model is a well-researched field in regression tasks. In linear models, the output is a linear combination of the input. In addition, it is noted that fully connected artificial neural networks (ANNs) “have been shown to reflect structure and function of the visual processing pathway”.2 As such, aside from linear models, ANNs is the other family of models that will be explored in this project.

3 Method

An effort was made to evaluate the “baseline” performance of the fwRF models used in NeuroGen. The performance is assessed on two main aspects: efficiency and accuracy. These two performance metrics would be measured simultaneously as the models are used to complete the following task: predict the average brain response to the 907 shared images in the NSD Dataset for any one of the 28 regions of the brain. The efficiency is simply measured by the real time (rather than the CPU time) it takes to finish this task, while the accuracy is measured by calculating the Pearson correlation between the model predictions with the actual average regional activations for the 907 shared images. For the most accurate model out of all the models that are created in this project, this same process would be used to evaluate its efficiency and accuracy. All tests are conducted on the same machine (a 2019, 15-inch MacBook Pro) using solely the CPU in an attempt to prevent machine’s load and technical specifications from interfering with the test results (particular the efficiency portion).

To choose the ideal hyperparameters, k-fold validation is used, as there are only 907 image-response pairs available for training. One single Python method was written to train and test the ANNs and the linear models. This ensures the consistency of the training process and the performance evaluation of the different models. The method was written in a way such that one can easily specify the model type and the number of folds in k-fold validation. Furthermore, the method also allows specification for the optimizing algorithm (e.g. stochastic gradient descent or Adam), the learning rate, and the epoch. There is also support for specifying the loss function (mean-squared error or negative Pearson correlation) that we are using the optimize our model.

Using this framework, various ANNs and linear models are tested in conjunction with different optimizers (stochastic gradient descent and Adam) and learning rates (7 log-spaced values from 0.00001 to 0.1). Since ANNs and linear models by nature cannot maintain spatial features like CNNs, the input to all the ANNs and linear models are based on the flattened feature maps (a vector of 9216 elements) after the fifth convolutional layers in AlexNet.

There are three ANNs with different structures that are explored in the project. The first ANN (ANN-1) has one hidden layer of 256 neurons, and an input and output layer of 9216 and 1 neurons, respectively. The second ANN (ANN-2) has two hidden layers, one with 256 and another with 128 neurons. The input and output layers are the same as ANN-1. The last ANN (ANN-A) has one input layer of 9244 neurons, two hidden layers of 512 and 128 neurons respectively, and an output layer of 1 neuron.

Note that the input layer of ANN-A has 28 extra neurons in the input layer than the other ANNs. This is because these 28 extra neurons form a one-hot encoding for the specific brain region for which the model should predict the average activation. As such, the ANN-A is designed with the goal that it can handle all possible regions in just one model. ANN-1 and ANN-2 can thus only predict the brain response of one particular region, and a separate ANN would need to be used if the activation of another region were to be predicted. Unfortunately, it is communicated by Zijin Gu that having one single model for all regions is not desirable. As such, there will be no further discussion regarding ANN-A. The other two ANNs will be discussed in later parts of this paper.

There are three linear regression models that are tested in the project. They only differ in the power that the flattened feature map of 9216 elements is raised to. In the first model, the input is formed by the 9216 elements unchanged (raised to the first power). In the second model, the input is formed by the original 9216 elements concatenated with them being raised to the second power. In the third model, an additional 9216 elements raised to the third power is concatenated.

All the models are optimized with the loss function being mean squared error. Although the training framework supports negative Pearson Correlation as a cost function to be minimized, this turns out to be ineffective as the model weights become extremely big. This is because Pearson Correlation can approach the maximal value of 1 as long as the model prediction and the actual average brain response forms a linear relationship. In other words, the correlation can be 1 even if the prediction deviates from the actual average activation significantly. There would need to be some regularization for the approach of optimizing using correlation to work, but that has not been done due to time constraints.

4 Results and Discussion

The accuracy of all the models are plotted in Figure 1. The ANNs trained with the ideal hyperparameters (Adam optimization algorithm with learning rate of 0.00001) has a consistently lower accuracy when compared to the fwRF model. However, ANN-2 is consistently more accurate than ANN-1: ANN-2 obtains a Pearson correlation that is around 0.1 lower than that achieved by the fwRF in all brain region while the ANN-1 is around 0.2 lower. Both ANNs seem to be capable of getting very high training correlation (extremely close to 1) and very low training MSE (extremely close to 0). However, the gap between the testing Pearson correlation indicates that such high training correlation and low training MSE is not an indicator of overfitting, as the more complex model (ANN-2, the one with two hidden layers) achieves a higher Pearson correlation than the less complex one.

The linear models performed significantly worse than the ANNs. As the power to which the 9216 input elements are raised to increase, the model performs worse. Even the best linear model has a Pearson correlation that is around 0.4 lower than the fwRF.

Chart, bar chart

Description automatically generated

Figure 1. The prediction accuracy for all the models across all brain regions. The y-axis is the Pearson Correlation. Note that for amygdala and hippocampus the fwRF model do not have values. This is because the fwRF models are not trained to predict the response in those regions. The “Regression-1” model is the linear model to which the input vector is raised to the power of 1. The same naming scheme applies to the other regression models.

The efficiency of the most accurate model (ANN-2) is measured and compared against that of the fwRF model (shown in Table 1). It is shown that despite having a Pearson correlation around 0.1 lower than fwRF model for all regions, the ANN can finish the task (predicting the average activation for all 907 images in one region of the brain) in less than 10% of the time it takes for the fwRF to do the same. This is because the fwRF model must load a subject-specific model, pass 907 images through it to get the subject-specific prediction, and repeat this process for all subjects before averaging the results. In short, using the fwRF model requires predicting the activation to 907 images 8 times while ANN only needs to do it once. However, it is worth noting that since the fwRF model predicts the brain response to the image for all regions of the brain at once, if the fwRF and ANN were tasked to generate the average activation for all 28 regions of the brain (instead of just one) to the 907 images, the fwRF model would still only need to pass the 907 images through the model 8 times, while the ANN model would need to pass the 907 images through the model 28 times, once for each region-specific model. In this regard, the ANN and the fwRF models should be used according to the specific context.

Table 1. The mean time in seconds across 3 trials for the fwRF model and ANN-2 to predict the average brain activation in a particular region for 907 images. The rightmost column is the percentage of time ANN-2 saves relative to the fwRF model.

5 Conclusion

It has been shown through this process that ANNs are a competent technique in this task of predicting the average activation of specific brain region to visual stimuli, although it still has a noticeable gap in terms of accuracy compared to state-of-the-art technology such as the fwRF models. It is also clear that linear models are not suitable for this task, as the accuracy is extremely poor even in the best-performing models. The best ANNs tested through the duration of this project sacrifices the high accuracy achieved by NeuroGen’s fwRF for a near tenfold improvement in the time it takes to predict one single brain region’s activation to an image.

To develop this project further, one can explore using the feature maps generated by other CNNs. Since AlexNet is relatively old (it’s made around 2012), newer neural networks may provide more informative feature maps. Furthermore, one can explore how the rest of the NSD Dataset can be used to better the average encoding model, as the current model only uses the shared set of 907 images. The vast majority of the dataset (the subjects’ response to their unique set of ~9000 distinct images) is still unused. Although this data is “incomplete” in a way that it is not the subjects’ average response to the image, it should still be of some value not only because of the sheer amount of data, but also because one subject’s activation to the image is a sizeable contribution (12.5% to be exact) to the average activation. In short, using feature maps from a more advanced CNN in conjunction with utilizing the rest of the NSD Dataset are directions worth looking into to further advance this study.

1. E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, L. T. Dowdle, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. Kal, “A massive 7T fMRI dataset to bridge cognitive and computational neuroscience”. <https://doi.org/10.1101/2021.02.22.432340> [↑](#endnote-ref-1)
2. Z. Gu, K. W. Jamison, M. Khosla, E. J. Allen, Y. Wu, T. Naselaris, K. Kay, M. R. Sabuncu, and A. Kuceyeski, “NeuroGen: activation optimized image synthesis for discovery neuroscience”. <https://arxiv.org/abs/2105.07140> [↑](#endnote-ref-2)
3. A. Amini. “MIT 6.S191 (2020): Convolutional Neural Networks”. [https://www.youtube.com/watch?v=iaSUYvmCekI](https://www.youtube.com/watch?v=iaSUYvmCekI&t=1833s) [↑](#endnote-ref-3)