CS230 - Group Project Final Report
Roman Pinchuk and Will Ross

Abstract

Human kind has continuously attempted to both understand and model the brain. These domains, Neuroscience and Artificial Intelligence, have in recent times, with the advent of Deep Learning, come together as elements of their respective insights begin to cross-pollinate. One area where this cross-pollination has not yet fully taken hold is the use of predictive neural networks to derive suspected relationships between different regions of the brain in forming semantic representations of visual stimuli. This study sought to leverage Shapley Value to interpret high-dimensionality models trained to predict what stimulus was shown to a human given a dense input of 5033 electrode readings. Ultimately, an ensemble of ten binary classification models revealed unique combinations of activity across signals detected in the Amygdala, Hippocampus, Parahippocampus, and Entorhinal Cortex. While these results are not in and of themselves affirmative discoveries of novel patterns in the semantic encoding processes of the brain, they do pave the way for new work in using interpretability of neural networks to decode brain signals.

Introduction

Historical studies evaluating single neuron electrode and scalpel EEG data have sought to demonstrate brain activity in the amygdala, hippocampus, parahippocampus, and entorhinal cortex in response to various stimuli, ranging from shapes and colors to happy and sad faces to natural and unnatural objects. Most of these activities have been undertaken with one of two objects: either increase human kinds' understanding of brain function or use that improved understanding to advance our ability to interact with technology through so-called "Brain Machine Interfaces" (BMIs).

Studies in this domain have traditionally been set up to measure activity in a suspect region of the brain in response to certain stimuli. For example, a human is shown an image of a sad face and output recordings from either scalpel EEG or single neuron electrodes are used to measure signal in the suspected regions. These signals can then be used in reverse to build and assess statistically learned models out-of-sample predictive ability.

To date, much of this work has been done with less sophisticated modeling techniques, such as Support Vector Machines, being applied to singular areas of brain activity (e.g., Amygdala vs. Hippocampus). By continuing on the work of Mormann et al.[1] and applying high-dimensional, multi-layer neural networks, we aimed to evaluate the predictive power of single neuron studies when evaluating all brain regions simultaneously. We then hoped to apply the work of Lundeberg et al.[2] to a new field, extracting Shapley Values across a wide range of electrodes inputs and using them to assess how different regions of the brain may "work together" in assessing a variety of different semantic classes.

Literature Review

The work relevant to this project sits at the intersection of four key areas of research -- Neuroscience, Brain Machine Interfaces, Deep Learning, and Model Agnostic Machine Learning Interpretability.

The state of the art in Neuroscience holds that semantic representations of visual stimuli, such as those presented in this study, are held as abstract representations in the brain.[3] Some progress has been made in the mapping of neurological representation of semantic concepts into categories such as living–nonliving or abstract–concrete.[4] Investigating semantic representations, in more detail in areas known to be activated in these events-- e.g., amygdala-- has been notoriously difficult. Mormann et al. collected a unique dataset that allowed four distinct regions of the brain -- amygdala, hippocampus, parahippocampus, and entorhinal cortex -- to be evaluated in their response relative to a group of 100 visual stimuli grouped in 10 semantic categories. Within each brain region, the study found varying predictive ability through the use of SVMs with 70-95% accuracy at the semantic class level (e.g., Bird) depending on the portion of the brain evaluated. Because each region of the brain was evaluated independently, it was difficult to infer relationships between areas of the brain from these models.[1]

In parallel to these developments in Neuroscience, Deep Learning has become a subject of great attention to the public eye due to the high level of generalization of it's models.[5] One noted criticism of this generalization, however, is that it comes at the inherent trade off of interpretability--that is, explanation of why a prediction is made relative to a set of input features -- when compared deterministic or lower dimensionality statistical models.[6] In an effort to advance the state of the art of interpretability, Shapley Values, popularized by Lunderber et al. and the related SHAP (SHapley Additive exPlantions) have emerged as a relatively resilient method for determining each feature of a machine learned model's contribution to a predicted value in terms of both direction and magnitude.[2]

Prior to this study, we are not aware of work that sought to relate state of the art Deep Learning and related Model Agnostic Interpretability methods to this field of neuroscience. It would seem that by using the entire feature space of electrodes to predict visual stimuli and then grouping Shapley Values by region of the brain, new insights to the field of Neuroscience could be extracted.

Dataset and Features

Our work continued on the results of Mormann et al. in "Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe".[1] We obtained a data set of 1000 observations (m=1000) across 100 unique stimuli (birds, computers, etc.) for 25 patients and corresponding single neuron readings from 5033 individual neurons (n=5033).

Stimuli were grouped into 10 semantic categories of 10 examples each. For clarity there were 10 unique birds, 10 unique items of clothing, 10 unique computers, etc. This dataset allows for classification on three different levels of granularity natural vs. man-made (ie, 5 semantic categories each - flower/bird vs. computer/clothing), the semantic class level (ie, flower vs. bird vs. computer vs. clothing), and unique image (ie, a shoe and a skirt both in the semantic class clothing).

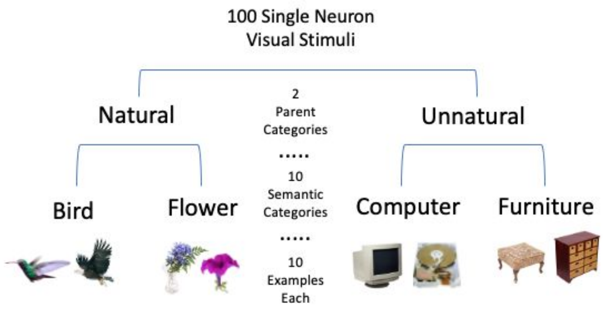


Figure 1: Hierarchy of Single Neuron Data Set
100 Images, 2 Parent Categories, 10 Semantic Categories, 10 Examples Each

Single unit neuron readings were likewise grouped into the parent categories of amygdala, hippocampus, entorhinal cortex, and parahippocampal cortex. Unfortunately raw sensor data was not available, so instead prediction on Z-scores comparing firing rates to a baseline, which the study obtained using Wave_Clus and Combinato, two common spike-detection algorithms that make use of convolutional masks to detect spikes in output readings.

Semantic Class	Mean Z-Value	Min Z-Value	Max Z-Value
Wild Animals	-0.0029	-3.5385	52.936
Fruit	0.0123	-3.5354	53.157
Flowers	0.0007	-3.3412	63.6854
Insects	0.004	-3.6217	35.4503
Birds	-0.0043	-3.2924	39.108
Manmade Food	0.0243	-3.5354	52.936
Clothes	0.0166	-3.383	41.1326
Furniture	0.0125	-3.2924	45.1599
Instruments	0.0044	-2.9275	62.2511
Computer	0.0075	-3.0495	52.936

Figure 2: Descriptive Statistics of Single Neuron Dataset

Methods

In order to establish an early baseline with our Single Neuron data set, we elected to use a simple feed-forward neural net built with Keras/Tensorflow. Early iterations indicated that variance/overfitting was a problem but our suspicion was also that the dimensionality of our feature space would require a network with high depth. We built a five layer multi-class feed-forward neural network (X=5033x1000 ||| 1024 relu, .6 dropout ||| 512 relu, .6 dropout ||| 256 relu, .6 dropout ||| 128 relu, .4 dropout ||| 100/10 softmax). As indicated, we experiment with the output layer as 100 unique classes vs. 10 unique classes in line with the hierarchal design of the original experiment. Due to our small sample size, we used only 10% of the sample as a holdout test set. We also elected to use the Adam optimizer and a mini-batch size of 100 due to some compute constraints. We implemented early stopping at 100 epochs for the 100 class model and 65 epochs for the 10 class model.

Given the high dimensionality of this task relative to the number of training examples, some of our primary concerns are overfitting and low accuracy. As we found in our initial experiments, these issues can be overcome in a variety of ways via regularization techniques, dropout, and early stopping to minimize variance/overfitting and upsampling to the semantic category level to improve accuracy.

Once positive results were obtained and optimized through a thorough search of the hyperparameter space, the same model architecture was broken into 10 individual binary classifiers with the only variation in the architecture being a change to the output unit. This ensemble of sigmoid classifiers had better compatibility with Shapley Values and also exhibited higher predictive accuracy in most semantic classes.

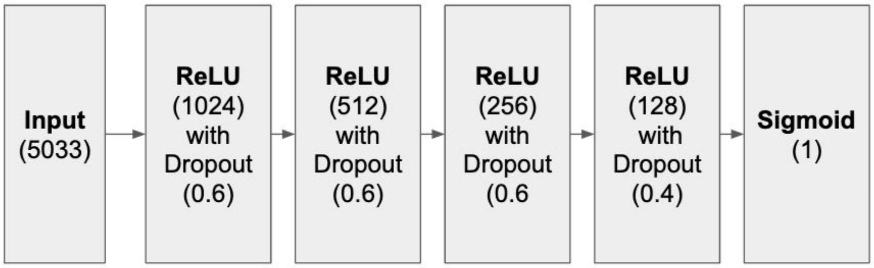


Figure 3: Final Five Layer Neural Network Architecture for Binary Classifier
Binary semantic classifiers were trained on all 5033 features for each of 10 classes

Lastly, we used the SHAP package Deep Explainer algorithm to evaluate these models. DeepSHAP uses regression to approximate Shapley Values for Deep Learning models. The approximation uses a distribution of background samples to estimate the relative contribution of each feature to the output of a model. By plotting Shapley Values for each feature in each observation against those of the dataset as a whole, we are able to determine the relative contribution of that feature to a given prediction. By summing these contributions across brain regions, then, we can determine the relative role our Neural Network suggests each region might contribute as follows.

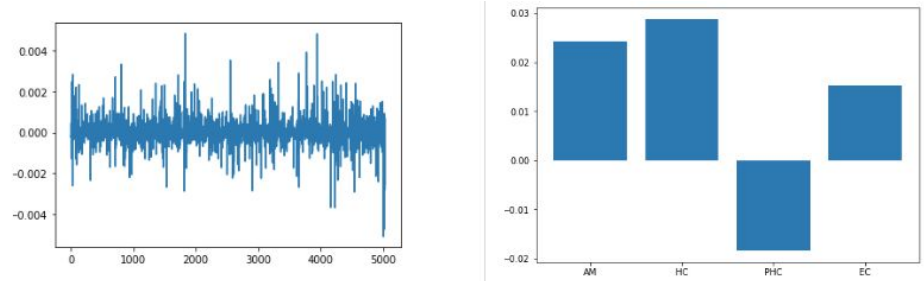


Figure 4: Example Shapley Values for 5033 Features and Corresponding Regional Distribution
High contribution of Amygdala (AM) and Hippocampus (HC) vs. mild contribution of Entorhinal Cortex (EC) and negative contribution of Parahippocampus (PHC) for predictions in class “Insects”

Experiments/Results/Discussion

After running our 100 class model (each bird, each flower, etc.), it became clear that the single neuron reading data simply did not have enough training examples to obtain meaningful results with out-of-sample accuracy of only 22% compared to train fit accuracy of 90%. We avoided this risk of high variance by simply upsampling the data using the same NN architecture across only 10 classes and saw significant improvement with out-of-sample accuracy of 95% and train fit accuracy of 98%.
These were astounding results as they indicated the validity of a neural network based approach and, in fact, already represented an improvement on the original paper which focused on support vector machines and achieved accuracy of only 70-95%, depending on the brain region analyzed.

The combination of our and prior results seem to indicate non-linear relationships between different segments of the brain. As a result, we used Shapley Values for interpretability on 10 distinct binary classifiers and saw high accuracy with all classes above 91% accuracy and an average of 97% accuracy across the dataset.

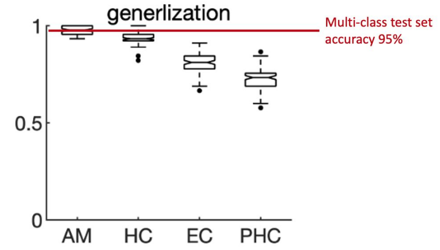
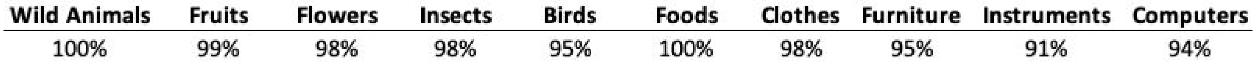


Figure 5: Comparison of SVM (Mormann et al.) vs. Multi-Class NN Predictive Accuracy
Box plots show accuracy across 100 cross-validation samples for the SVM approach
AM = amygdala, HC = hippocampus, EC = entorhinal cortex, & PHC = parahippocampal cortex



Wild Animals	Fruits	Flowers	Insects	Birds	Foods	Clothes	Furniture	Instruments	Computers
100%	99%	98%	98%	95%	100%	98%	95%	91%	94%

Figure 6: Accuracy of Binary Classifiers for Each of Ten Semantic Classes After 100 Epochs

Beyond this high level of accuracy, we were particularly pleased to see that Shapley Values did show distinction between the relative contributions of electrodes in each brain region to out-of-sample predictive accuracy. For example, while the Amygdala was the leading contributor in most classes, the Wild Animals, Insects, and Instrument classes all saw high contributions from the Hippocampus. Particularly compelling was the indication that the presence of stimuli in the Parahippocampus actually had a negative impact for classification of Insects.

Conclusion and Future Work

We believe the variety of methods and datasets explored and results achieved in this paper represent progress in the fields of both core Neuroscience as well as the application of Deep Learning to Brain Machine Interfaces (BMIs). Evaluating a variety of different brain stimuli at multiple levels of specificity and applying interpretability methods to these results, we have demonstrated the pros/cons of different approaches and also pathed the path for a variety of feature avenues of work.

With regards to Single Neuron Data, we showed that these datasets, while unrealistic for BMI, have incredibly high predictive accuracy, especially when all regions of the brian are examined in parallel. This represents significant progress from prior studies, which looked only at single region activity.

Last but not least, through the use of Shapley Values, we were able to indicate that different semantic classes are optimally represented by predictive Neural Networks that place different levels of feature attribution on the Amygdala vs. Hippocampus vs. Parahippocampus vs. Entorhinal Cortex.

While this work cannot guarantee that these brain regions are, in fact, utilized differently in the encodings of these specific semantic representations, it does present the opportunity for future work to cross-validate this finding via other means. While it remains true that Neural Networks are not, in fact, accurate representations of the human brain, we are optimistic that the combination of these types of models and the advances of interpretability of them will help to advance the fields of neuroscience and Brain Machine Interfacing for many years to come.

Contributions

Our team attempted to work collaboratively across all phases of the project. We both contributed to the relevant neural network architectures, results analysis, and subsequent “grad student descent”.

We’d also like to thank Dr. Ueli Rutishauser of California Institute of Technology and Cedars Sinai for directing us to the Single Neuron data set and suggesting to us that interpretability of a multi-region model would be of genuine help to advance the field.

References

<1> Reber TP, Bausch M, Mackay S, Boström J, Elger CE, Mormann F (2019) Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe. PLoS Biol 17(6): e3000290. https://doi.org/10.1371/journal.pbio.3000290

<2>Scott Lundberg, Su-In Lee: A unified approach to interpreting model predictions. CoRR abs/1705.07874 (2017)

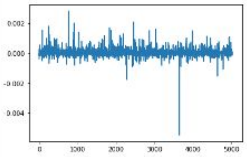
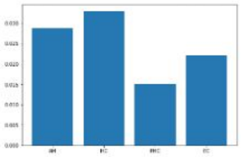
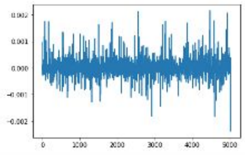
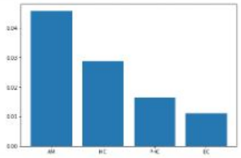
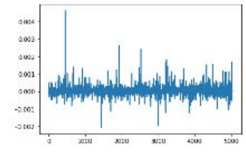
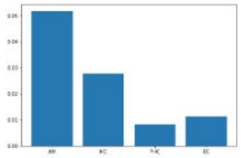
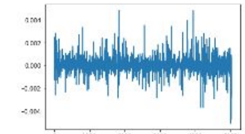
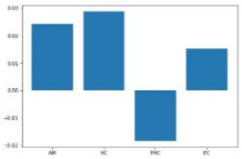
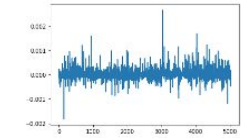
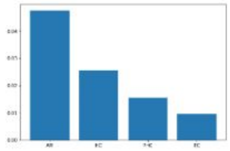
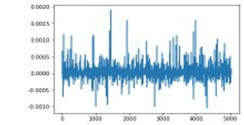
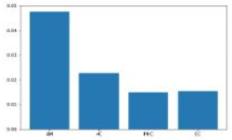
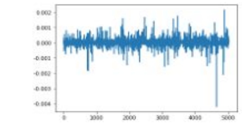
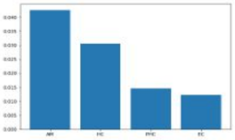
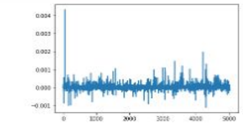
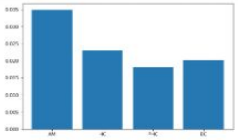
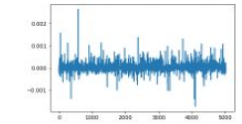
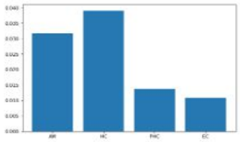
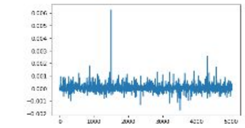
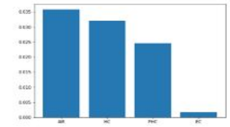
<3>Patterson K, Nestor PJ, Rogers TT. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci. 2007;8: 976–987. Pmid:18026167

<4>Yee E, Jones MN, McRae K. Semantic Memory. In: Wixted JT, Thompson-Schill SL, editors. The Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience. 4th ed. New York: Wiley; 2017.

<5>LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* 521, 436–444 (2015). https://doi.org/10.1038/nature14539

<6>Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

Appendix A - Summary of All Results

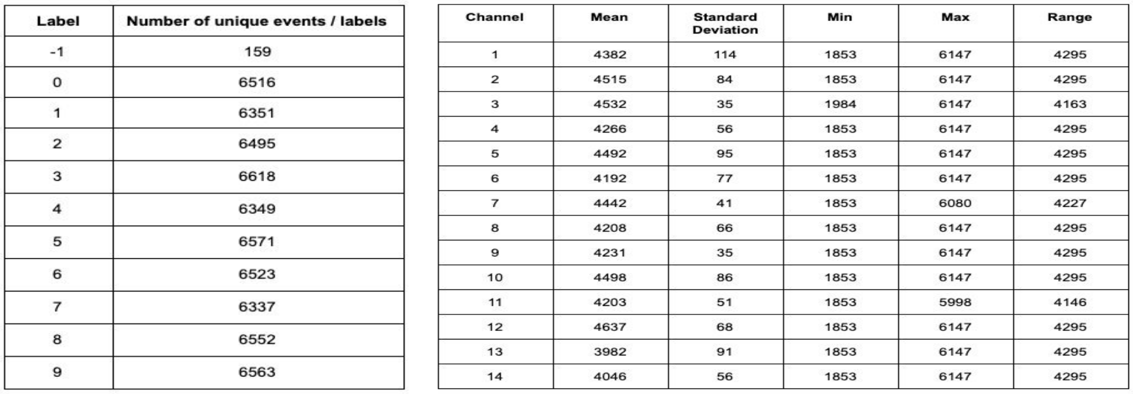
Parent Category (Prior Study Task)	Semantic Category Name (10 Unique Stimuli Per Class)	Example Stimuli (n=10x10) (note - Human Exposed, not model)	Classifier Test Accuracy (input - Electrode , ouput - Class, m=1000)	Shap Value Plot - Positive Example (interprets most 'predictive' electrodes in model)	Regional Stimulation (use Shap Values to evaluate regional stimulation)
Natural	Wild Animals		100%		
Natural	Fruits		99%		
Natural	Flowers		98%		
Natural	Insects		98%		
Natural	Birds		95%		
Unnatural	Manmade Foods		100%		
Unnatural	Clothes		98%		
Unnatural	Furniture		95%		
Unnatural	Instruments		91%		
Unnatural	Computers		94%		

Appendix B - EEG Research - Alternative Dataset

Due to the interdisciplinary nature of our research, we evaluated two different datasets in attempting to determine the applicability of Deep Learning and Shapley Values to neuroscience research. We ultimately found Single Neuron Readings to outperform the EEG data.[1]

EEG readings of brain waves from the MindBigData Dataset

Initial research into EEG readings was done using a publicly available dataset, MindBigData, of brain wave readings collected using the Emotive EPOC device.[1] This dataset evaluated responses from a single human subject being exposed to digits from 0-9 and, as a control, no digit at all. This data offered limited variety in individuals observed and was limited to only numeric characters. It's advantages however came in that it offered 65034 unique observations (m=65034) and provided access to raw time series data over two seconds recorded at 128Hz (Tx=320).



Label	Number of unique events / labels
-1	159
0	6516
1	6351
2	6495
3	6618
4	6349
5	6571
6	6523
7	6337
8	6552
9	6563

Channel	Mean	Standard Deviation	Min	Max	Range
1	4382	114	1853	6147	4295
2	4515	84	1853	6147	4295
3	4532	35	1984	6147	4163
4	4266	56	1853	6147	4295
5	4492	95	1853	6147	4295
6	4192	77	1853	6147	4295
7	4442	41	1853	6080	4227
8	4208	66	1853	6147	4295
9	4231	35	1853	6147	4295
10	4498	86	1853	6147	4295
11	4203	51	1853	5998	4146
12	4637	68	1853	6147	4295
13	3982	91	1853	6147	4295
14	4046	56	1853	6147	4295

Figure 1: Distribution of Labels and Data Statistics by Channel for MindBig Dataset

As the above table shows, the readings across the 14 channels for the EEG data set were very consistent in terms of the min and max values, likely reflecting both the biology and limits of the device itself, as well as in terms of the mean signal value. However, there appears to be some difference in terms of standard deviation between channels.

EEG readings of brain waves from the MindBigData Dataset

Due to the time-series nature of EEG data, we initiated our research by experimenting with an LSTM neural network architecture, implemented using Tensorflow/Keras. We used 128 LSTM units, with an input shape of (320, 14), corresponding to the max number of readings per channel (320) and the number of channels (14).The second layer had 512 ReLU units, and the final layer was a Softmax with 11 units, corresponding to the 11 possible classes. We used the sparse categorical cross entropy as our loss function and the ADAM optimizer. We used a training/test split of 90/10 and trained the model for 100 epochs.

Unfortunately, our accuracy barely exceeded 10% for the test set, and so we did not focus on regularization/dropout at this stage. After failing to achieve satisfactory results even after trying multiple publicly available EEG datasets and experimenting with various RNN architectures we decided to focus our efforts on Interpretability of the Single Neuron Readings. We hope that future work in this area will achieve better results, possibly by moving away from using RNN in favor of CNN, as this kind of research holds substantial theoretical and practical importance to building Brain-Machine Interfaces.

[1] Vivancos, D. This MindBigData The "MNIST" of Brain Digits http://www.mindbigdata.com/opendb/