Shateesh Bhugwansing

CSC 59867 | Senior Design II

Prof. Michael Grossberg

21 December, 2018

Final Report: Analysis of EEG characteristics in Identifying Congruent Modal/Lexical Stimuli

Abstract

The Stroop Effect is a popular area of interest in Psychology research because it raises many questions about the role of language and non-language stimuli in selective attention. One particular study examined the entire distractor selection paradigm, including all modality and lexicality combinations of both target and flanker stimuli, in order to identify the configuration of factors that affect distraction to both the greatest and least extremes. The aforementioned study analyzed the Event-Related Potential (ERP) response taken from electroencephalography (EEG) measurements of the test subjects.

The project discussed in this paper aims to provide an automated method of identifying and visualizing such responses using machine learning (ML). The machine learning algorithms used in this project were set up to classify EEG data that were labeled in terms of their modality, lexicality and semantic congruence. The decision boundaries formed by the learning models were then mapped to electrode positions around the brain to show where and when in the brain a classification decision was most influenced by. This project resulted in a fundamental pipeline of methods that can be used, but there is much work to be done in terms of tuning the parameters to get more dependable results.

Introduction

Interference or inhibition have been largely studied since early as 1890 and has been a large part of scientific research. The famous “Stroop Effect” named after researcher J. Ridley Stroop attributed the powerful effects of interference of color word stimuli upon visually seeing colors. Reading the color red written in blue color causes interference. We are receiving two non-congruent stimuli from a sensory (visual) point and reading a word (lexical). Picture-word analogue of the stroop effect has also been researched. In the picture-word interference (PWI) task, subjects name objects in picture while having a distractor word written on the picture. The distractor can be the name of the picture (pictured cat, word cat), categorically related word (pictured cat, word dog) or categorically unrelated word (pictured cat, word pen). Research findings on PWI is similar to the findings of the stroop effect. Trials with categorically related distractors such as a picture of a cat and word dog(category: animal) slow down response times opposed to semantically congruent picture and word such as picture of cat and word car. This finding is analogous to that in the colour word

Stroop task. The different color word and font color slowed response times(category:color) opposed to color words all written in black.

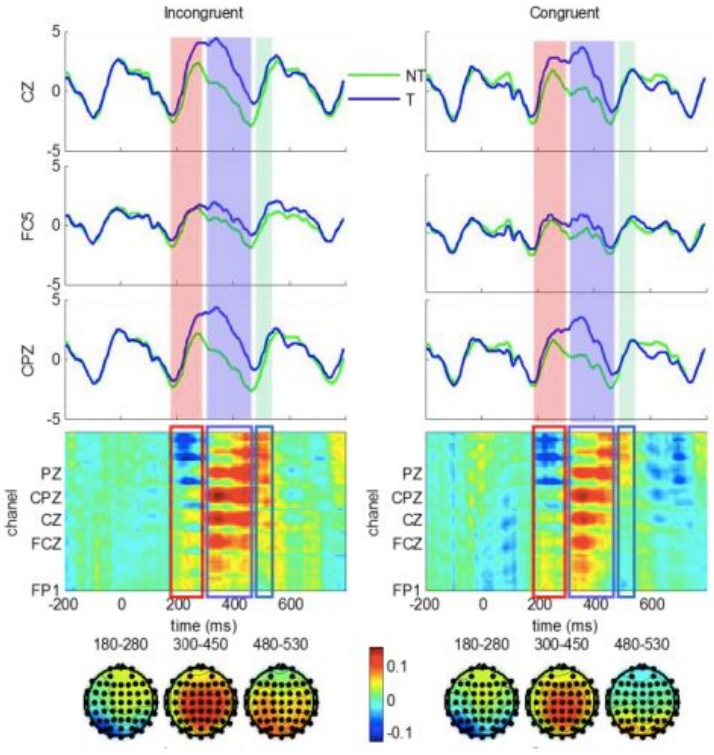


Figure 1: Readings shown between 200 ms and 500 ms where the highest class discrimination was noticed. The radical change in color from neutral green shows that either ERP target amplitude was higher than ERP distractor or vice versa.

Recent research has introduced ways to reduce or even reverse the effects of an incongruent word. Dave Britton from the Graduate Faculty of Psychology, the City University of New York, documented brain neural activity using EEG, response times, and accuracy of selecting correct target stimuli [1]. The flanker selective paradigm used 20 different combinations of visual/auditory modalities with word/nonword lexicalities as both flankers (distractors) and as targets to manipulate attention with phonological congruent and non-congruent trials. Our goal as data scientists is to help discover characteristics of EEG that help identify where in the brain semantically congruent stimuli are recognized. We want learn where and how semantic content is stored in the brain. Thus, we aim to discover the parts of the brain that work together and investigate the signal outputs of the brain given, semantic congruent stimuli(distractor/target).

We investigated the EEG signals using common machine learning algorithms that were taught in the introductory portion of this course. The dataset contained the trigger file used in the experiment where the EEG data was initially collected, which allowed for the data to be labeled and passed through supervised learning algorithms. The choice of algorithm depended on its availability in a Python library (e.g SciKit-Learn) and its performance in other academic studies that classified EEG using machine learning as well [2,3].

BACKGROUND

Integrated information from different modalities are crucial for information processing [1]. The convergence of information from multisensory input enhances behavioral performance (e.g, speed and accuracy) due to the increase of certain neurons activating (firing) together. The firing rate of cells of multisensory inputs far exceeds firing rate of unisensory inputs. This means that humans are capable of realizing kind of input and storing information when we have more than one sensory organ being used(such as seeing and hearing)Semantic Congruence refers to the combination of multisensory stimuli that are presented in terms of the same meaning. The impact of multisensory input has been investigated in several studies [2,3] and significantly faster times were found in semantically congruent audio-visual pairings compared with unisensory input. Significant longer times were found for mismatched incongruent audio-visual pairings. Thus, semantic congruence has an significant impact on the integration of information across different modalities.

Two visual-audio bimodal stimulus (VABS) systems were conducted on a study[4]. The study was comparing congruent and non-congruent VABS based model. The paper aimed to see if semantically congruent stimuli can get the same performance as incongruent stimuli in Brain computer interface (BCI) systems. The results showed higher amplitude of the Event Related Potential (ERP) of semantically incongruent non-target and target stimuli. The acquired raw EEG data was preprocessed for artifact removal through Independent component analysis(ICA). Data was than filtered through Band pass of 0.5-40Hz and down sampled to 200Hz. The event related potentials of three channels (target ERPs and distractor ERPs) were calculated and averaged for each condition (congruent and incongruent). They plotted the ERP of each channel (CZ, FC5 and CPz) for each condition (congruent and incongruent) with two colors representing the target stimuli and distractor stimuli. The spatial temporal distribution of class discrimination (target and distractor) were shown for both conditions. The scalp map was shown depicting the average sign value for three specific time intervals. A Positive value represents that target ERPs have larger amplitude than distractor ERPs. Whereas a negative value represents the opposite. Support Vector Machine (SVM) was used to do binary classification of target and non-target. The illustration of Fig 1 demonstrates that higher class discrimination values are lying at channels around FCz, Cz, CPz,

and Pz with the time interval from 200ms to 500ms approximately.

From this research we can learn to visualize the patterns associated with each target ERPs and distractors ERPs so that we can find evidence to differentiate characteristics of EEG. This will help us classify the type of information contained in the EEG (e.g visual, audio, language, non-language), and find sufficient evidence in the EEG that shows brain has understood semantic congruent stimuli was seen. We can see from figure 1 that there is no significant evidence to classify congruent and incongruent from the results because results are similar. The research explains that they might not have had enough data to get better results. We aim to make sure that our data is sufficient enough to generate better results.

Data

The purpose of collecting data from subjects is to see what happens in the brain when we present stimuli that are different from target stimuli. The experiment also measures response times and accuracy with flanker distractors vs target. There are 4 types of stimuli available (a picture, sound the thing makes, thing spelled visually, and the thing spoken), two of which are auditory/visual (modality) and two are language/non-language (lexical). Using the 4 types of stimuli, 20 combinations are used with either varying modality or lexicality. The instructions for the subject is click left or right to select appropriate target on screen. The time it takes to respond and accuracy is taken into consideration.

There were 32 subjects that each experienced 2 full runs of the experiment. Each run of the experiment produced 1,280 trials. Each trial consists of four stages: focus time, flanker, target, flanker. Each stage is presented for 500 milliseconds, resulting in a total of 2 second trials. Each stimuli time window is 500 ms, however not all stimuli is shown for the same time. Each stimuli is shown for a time within 500 ms depending on the sensory input. For example the word “DOG” is shown for 50ms, than black screen for 450ms. The sample rate to collect data was 512 samples/sec, which results in 256 readings for each stage. This experiment produced a total of 9.8GB of data, which includes all of the EEG readings and the triggers for each session.

Methods

The overall goal of this project is to attain the ability to identify congruence, or lack thereof between the flanker and the target stimuli from EEG data alone. In addition, we aim to analyze our machine learning algorithms to identify the deciding factors that affect the final result. We predict that these deciding factors will correlate with the parts of the brain that are making such decisions.

Before attempting any classification, the raw EEG dataset had to be preprocessed. Noise, such as ocular and muscular artifacts (spikes in brain activity

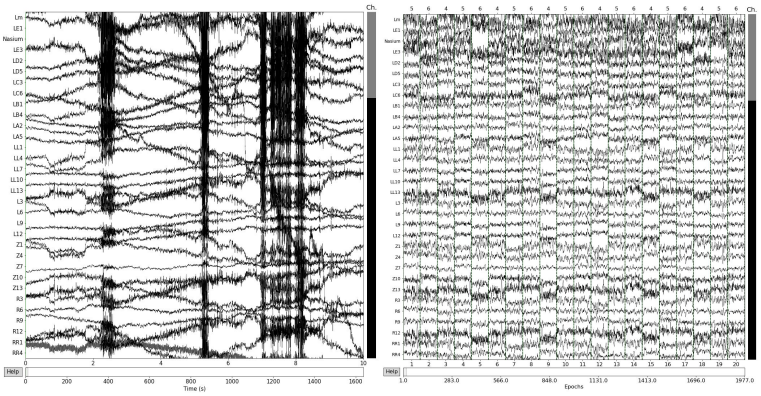


Figure 2: The raw EEG data before (left) and after (right) applying ICA for artifact correction

caused by movement of the eye or muscles in the subject’s body), needed to be filtered out. The python library MNE [9] provides a method to perform artifact correction using ICA. MNE has a built-in ICA object that can be fitted onto any raw data file.

After artifact removal, the raw data was then converted into 500 ms epochs, as defined in the experimental setup where the raw data came from [1]. In the original experiment,

subjects were exposed to stimuli for up to 500 ms before another one was presented. One complete trial consisted of three, 500ms intervals, followed by a rest interval of 500ms where no stimuli was presented, and the subject had to identify the target stimuli of that trial. The trigger files that accompanied the raw data included event IDs that described the entire trial, including the configuration of modality and lexicality that was presented. For the first couple of classification problems such as modality and lexicality (defined further along in this paper), the event IDs of interest were in the middle of the trial, and thus extracted in this preprocessing phase. Each stimulus had its own event ID (code), and thus a combination of three "stim codes" made up the content of one trial [1]. Figure 3 illustrates the events that our group wanted to obtain from the raw data.

After the data had been preprocessed, our team worked with Dave Britton, the Psychology expert whose data we're using, and Michael Grossberg, the Senior Design project advisor, to devise a list of classification problems that we could tackle. These problems gradually increase in difficulty, until eventually answering the question on congruence. In this context, "difficulty" refers to the ability to classify the EEG data with high accuracy. In addition, 'difficulty' refers to the ability to identify the classification model's coefficients with electrode positions on the brain. For each of these problems, we will explore the factors that the algorithms choose to base its decisions on and map them to the human brain, which allows psychologists to understand the part of the brain is most influential on a specific problem. The list of binary classification problems were proposed as follows:

1. **Identify the modality of the given stimuli (audio vs. visual stimulus)***
2. **Identify the lexicality of the given stimuli (language vs. non-language stimuli)***
3. Identify the semantic content of the stimuli (e.g dog, baby, etc.)
4. Can you tell the type of target? (auditory vs. visual AND language vs. non-language)
5. Given the EEG readings from an entire experimental trial, can you identify congruence between flanker and target stimuli?

*Note: * indicates the problems that we were able to build a pipeline for at the time this paper is being written.*

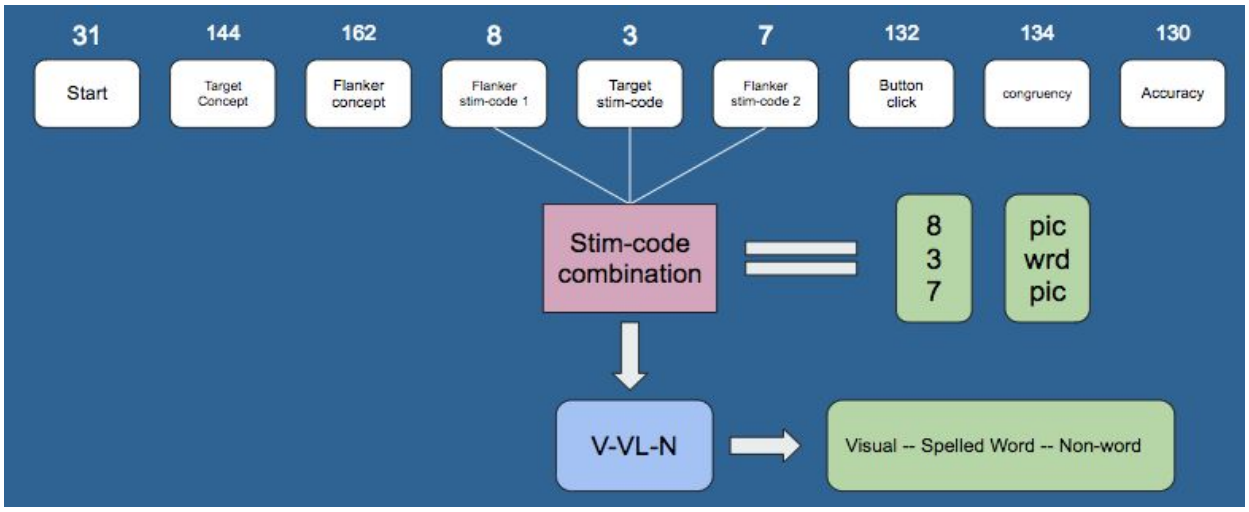


Figure 3: An illustration of the preprocessing step that extracts the events with the event IDs of the stimuli shown in a trial.

The modality question is considered easiest because there is a well defined region in the brain for where audio and visual signal processing happen. Questions about the semantic content of the target become more complicated because there are multiple factors that can be tested for, that may not be suitable for a binary classification model. At the moment this paper is being written, our team had built the pipeline for the first two problems in this list.

For each classification problem in the list above, multiple classification methods were used. Throughout the very early stages of the project, our team experimented with linear regression, random forests, linear discriminant analysis (LDA) and linear support vector machine (SVM). Although we started out with these four, as we progressed to the visualization stage of the project, we used Logistic Regression and Random Forests first. At the time of the creation of this paper, these two algorithms are the only two that were were able to visualize. Logistic Regression and Random Forests were the first algorithms chosen due to their simplicity in extracting their coefficients/importance matrices that were generated in the training process. If this project continues on, the same classification and visualization methods will be used on more algorithms and the performance will be compared.

My individual contribution involved using the Logistic Regression model to develop a comprehensive visualization technique that showed the electrode positions (channels) that were most useful during the classification process. The results were displayed in two ways; one was using built-in MNE functions, while I came up with another for this specific purpose.

The visualization that I came up with involves first creating a pipeline for classification that includes the following: 1) Vectorizer, 2) StandardScaler, 3) Logistic Regression Model. MNE's Vectorizer is used to transform the shape of the input data from 3 dimensions (number of epochs x number of channels x number of samples) into 2 dimensions (number of epochs x number of features). The feature dimension becomes a 1D combination of the temporal and spatial elements of the data, which is necessary for us to use the SciKit-Learn model for Logistic Regression. SciKit-Learn's StandardScaler removes the mean from the data and scales it to unit variance, which helps the features to fit a standard normal distribution. This pipeline is then fitted onto the epoch data and its labels. The labels are manually generated, based on some pattern in the events IDs. For example, in the case of the modality problem (audio vs. visual), all events less than 700 were audio signals, and greater than 700 were visual signals. In the case of lexicality, if the first digit is odd, the signal is language, and if its even, the signal is non-language. With the pipeline assembled, the next step was to cross validate the classification by using stratified K fold. 10 was the chosen number of folds, after trial and error with values between 1 - 15, and seeing the performance plateau after 10. After fitting the pipeline to the training set of data and labels, the coefficients are extracted by using the MNE built-in function "get_coef", which is a wrapper method around SciKit-Learn's "coef_" attribute that exists for its logistic regression model. Next, the classification is performed on the test data, and the scores are calculated using SciKit-Learn's built-in "Score" method. For each iteration of classification in the cross validation process, a new set of coefficients are generated each time, resulting in 10 matrices, each being 125x257. There are 125 rows, representing the number of channels that are being used (we had to drop three channels because those were reference nodes and did not provide any useful data), and 257 samples taken across the 500ms interval. For each matrix, the coefficients for each channel were averaged across all 257 samples, and then took the absolute values and created a python dictionary of each value to its corresponding channel name. This dictionary was sorted by value in descending order, which now ranks the most used channels in the binary classification

decision in logistic regression. Principal Component Analysis (PCA) on our data showed that 20 components is guaranteed to retain over 99% variance, so the top 20 channels were selected from this sorted list. The channels that appeared in this top 20 list for each iteration in the K folds process were then identified on the montage plot of the EEG cap used in this experiment [1].

While this method is effective in showing the most important channels to the logistic regression model, it averages out the time dimension, which is effectively losing information there. The second visualization technique is a remedy to that issue. The second method used in this project came from an example in the MNE documentation which uses the "SlidingEstimator" object, which fits and predicts a classification pipeline onto a series of subsets of data. In the context of this project, the sliding estimator fits the logistic regression pipeline, minus the vectorizer, onto each sample of an epoch. The vectorizer is not needed here because we actually need the third dimension (number of samples, 257) to pass the Sliding Estimator over. The estimator is scored using the same K fold cross validation, and then averaged across each fold. The scores are then plotted as an ROC AUC curve over time, showing the times when the rate of true positives are highest. This plot is useful in determining which time intervals are most often used in the classification process. This plot is paired with a plot of the evoked (averaged) coefficients values over time, highlighting what the coefficient patterns look like at "peak" times throughout the epoch. In this context, "peak" refers to the highest coefficient values generated by fitting the logistic regression pipeline onto the training data. The latter plot is useful to Psychology researchers because it's something very similar to what they see when visualizing ERP responses. The new visual here is the patterns on a topographical map of the EEG montage. MNE provides a parameter to generate these visuals for fixed time intervals, but we ultimately decided on displaying the patterns at the most active peaks for a better understanding of exactly when the algorithm finds the most useful information to perform classification.

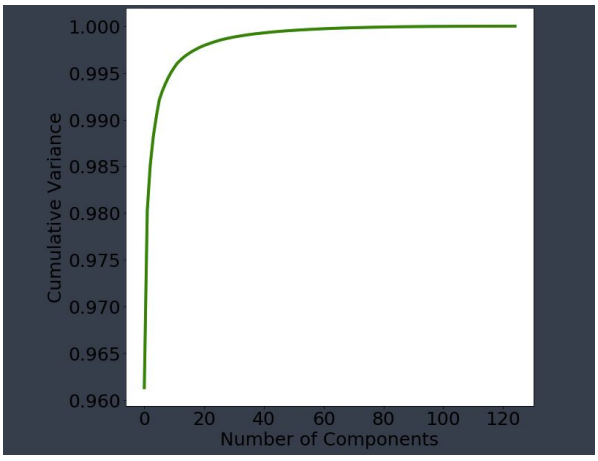


Figure 4: PCA on the EEG data shows that 20 components guarantees over 99.5% of variance is contained.

A key point to the methodology is that while the overall method of classifying and visualizing was being developed, the data size was limited to just one subject file at a time. This allowed us to minimize the run time of our experiments so we can edit our progress if necessary. Towards the end of the project, once these methods were agreed to by the rest of the team, the experiments were conducted on larger chunks of the entire data set. There were memory restrictions on our computing node provided by the City College of New York, limiting us to a maximum of ⅛ of the data set in one experiment.

Results

As previously mentioned, at the time of the creation of this paper the team was able to build a complete pipeline for the first two classification problems from the master list outlined in the Methods section. I was responsible for executing the complete pipeline for the modality and lexicality problems using Logistic Regression. These results will be compared with the results of the same methods but using Random Forests, that the other half of the team used. As expected, the Random

Forests performed better than Logistic Regression in terms of accuracy, as seen in Table 1. Also, in terms of visualizing the most important channels, the "importance" matrix generated from the Random Forest appeared more precise than Logistic Regression. The importance matrix highlighted a much more precise area of the brain being used the most, while the biggest Logistic Regression coefficients spanned across a much wider region.

Table 1. The average classification scores of both Random Forests and Logistic Regression, for the modality and lexicality problems.

Classifier	**Avg. Accuracy, Audio vs. Visual**	**Avg. Accuracy, Language vs. Non-Language**
Random Forests	69%	60%
Logistic Regression	63%	55%

Random Forests was always expected to perform better than Logistic Regression due to the fundamental design of the two algorithms. Logistic Regression does not perform well when the feature space is too large because each feature is assigned a weight (the coefficient matrix that we're targeting). If there are too many features, the weights start to become very similar to each other, which is exactly what happened to this experiment. The variance between coefficients was on the scale of 10^-6, about 100 times smaller than the largest coefficient. Random Forests, on the other hand, is a combination of decision trees that are independent of the size of the feature space. The combination of random decision trees helps to avoid overfitting, which improves the average performance across multiple folds in cross validation.

The accuracy scores that were achieved in our experiments is similar to other publications [2,3] that attempted to use machine learning techniques on EEG data. Generally speaking, machine learning with EEG is fickle because of the extremely high

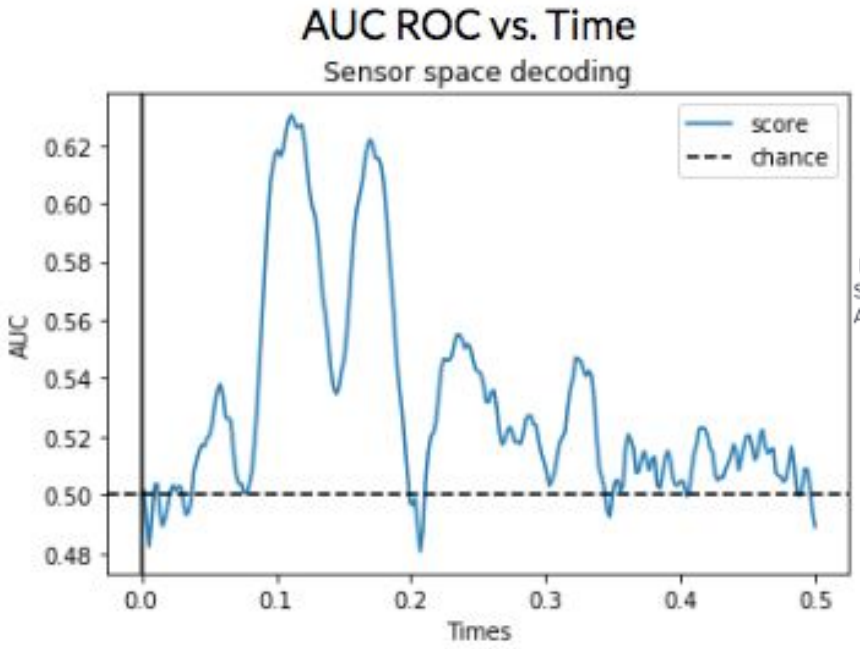
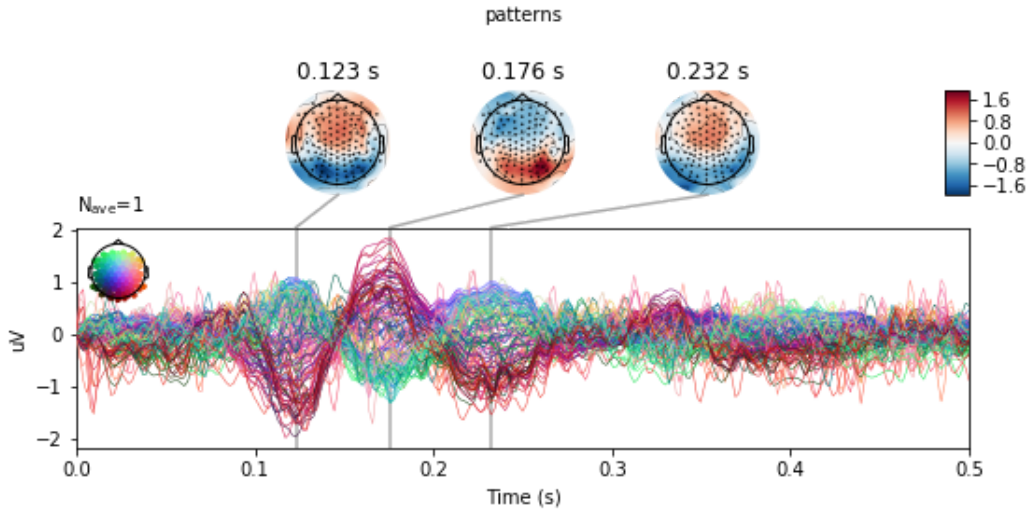


Figure 5: The AUC ROC Curve for the Sliding estimator visualization technique, Audio vs. Visual, using Logistic Regression.



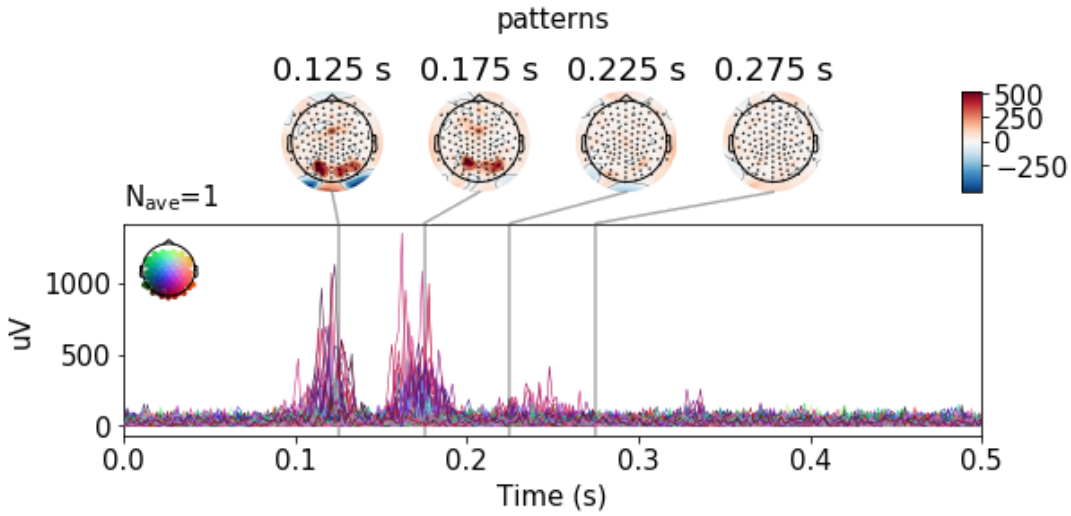


Figure 6 (top): The coefficients patterns plot over time for Audio vs. Visual, using Logistic Regression, with topographical maps at the times with highest variance.

Figure 7 (bottom): The importance patterns plot over time for Audio vs. Visual, using Random Forests, with topographical maps at the times with highest variance.

sensitivity of EEG data. When looking at individual data points, there is little consistency between the same test subject in different trials, let alone different test subjects. There are also a variety of ways to introduce noise to the recording. Even though there are toole like MNE that provide reusable functions to remove artifacts from raw EEG, there is no guarantee that *all* noise will be removed from *every* data sample. In addition, the classifiers that we used

are generally not the best at accurately classifying test data; other binary classifiers such as Linear SVM, along with Deep Learning networks, are well known for out performing Logistic Regression and Random Forests. However, as stated earlier, these two classifiers were chosen as a first step because of their ease to implement and extract importance features. In the future, more complex models can be implemented using the pipeline that we developed in this project.

Figures 5 and 6 illustrate the AUC ROC curve and the MNE topographical plot of the patterns in the Logistic Regression coefficients over time, for the modality problem. Be reminded that this visualization technique is the second one mentioned at the end of the Methods section, meaning that the results were generated via the Sliding Estimator. This means that the coefficients that Logistic Regression create are independent from each other at every time sample. We can see in figure 6, at t=0.176 ms, the colors of the coefficient pattern are practically inverted from the topographical map at t=0.123 ms, because the audio and visual signals are processed at different times during an epoch. Figure 5 shows the AUC ROC curve which confirms that the highest probability of scoring true positives in classification is between 120 ms to 230 ms, which is what Dave observed to be the most active time segment in the data from his original experiment. Figure 7 shows the same type of plot as in figure 6, but using Random Forests. Here, the difference in importance is much more drastic between channels, and only a select few are most obvious. These channels do correlate with the results from Logistic Regression in that they are located in what appears to be the visual cortex of the brain, which is commonly known to be in the back of the head. Figure 8 illustrates the channels with the highest coefficients using the the first method of visualization outlined in the Methods section of this paper. This plot is not consistent with the plots in figures 6 and 7. The regions highlighted in figure 8 are further forward, and further to the left and right than the regions that are more colorful in the topographical maps. Figure 8 is the reason why I decided to stop using this method towards the end of the project. Not only does it not capture the time element, but there seems to be an issue with how the dictionary was created with the channel names and the coefficients. There would need to be more investigation into this custom method for visualization, but as a team we decided to move forward with the MNE example, due to time constraints.

The lexicality problem proved to be more difficult to classify and identify the most used channels in the brain. Table 1 shows that the classification scores were lower in both classifiers. The psychology reasoning is that in the case of lexicality, there are more decisions being made in the brain before the decision of language or non-language. This idea is proved because in the lexicality problem, the most active regions appear a little later in time, between 130-300 ms. Therefore, the factors influencing lexicality are more complex than the factors influencing the modality problem. In the future, a stronger signal for the lexicality problem could be identified by limiting the data to one type of modality, e.g. classify for language vs. non-language on only all audio

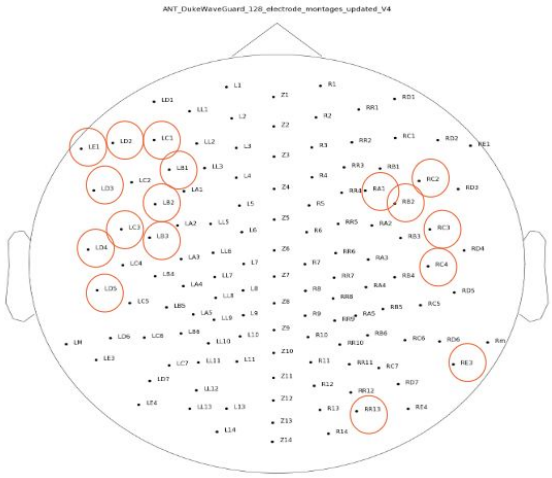


Figure 8: The channels with highest coefficient values in the Audio vs. Visual problem, using Logistic Regression.

stimuli. This future experiment would need to include a paradigm that tests for all possible

combinations of factors, starting with modality and then including factors such as semantic context and congruence.

Figures 10 and 11 show the same plot as figures 6 and 7, but instead for the lexicality problem. Unlike the modality problem, there is a stark difference in the location of important channels for Logistic Regression and Random Forests. Figure 9 shows that the significant features for logistic regression appear in the back of the head, similar to the pattern in the modality problem. However, this result may very well be due to poor classification performance. The AUC ROC curve in figure 9 shows that the rate of true positives never goes beyond 55%, which is only just above average. The performance of Random Forest in this case is 60%, which gives us reason to believe the results in figure 11 more. In figure 11, the channels of most importance to the Random Forest classifier is consistent between 130 ms and 300 ms, indicating that there may be a significant amount of activity in this region of the brain that correlates with language processing. Future work for this project includes confirming the lexicality result with Random Forests with the psychologist who provided the data set.

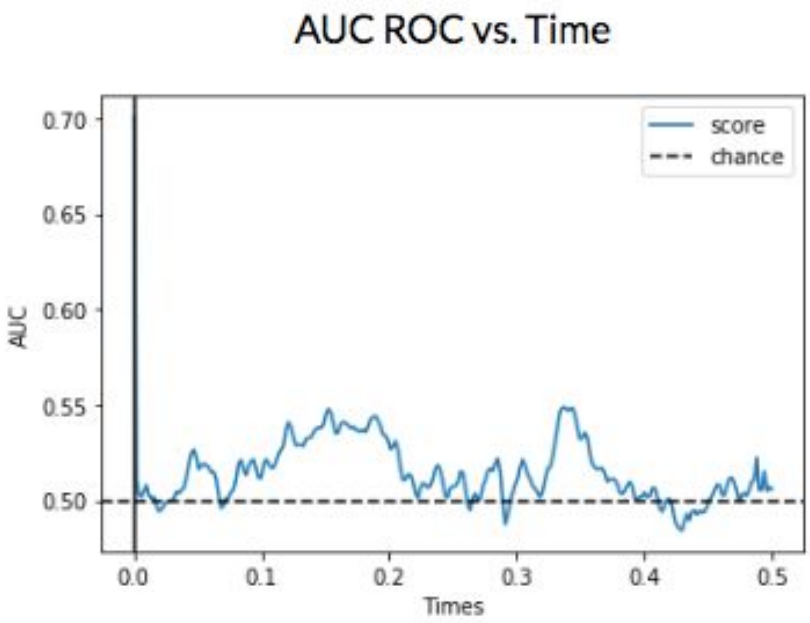
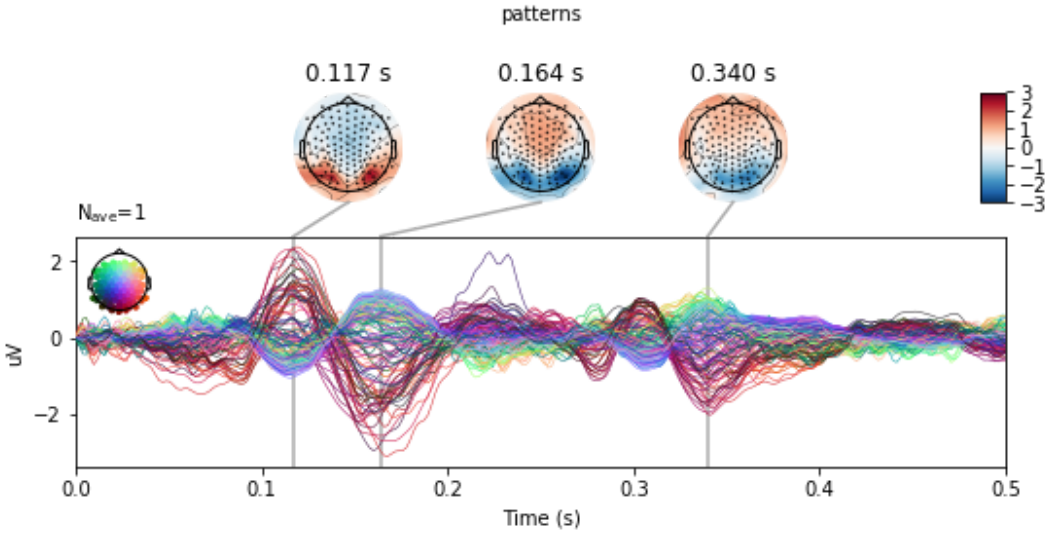


Figure 9: The AUC ROC Curve for the Sliding estimator visualization technique, Language vs. Non-Language, using Logistic Regression.



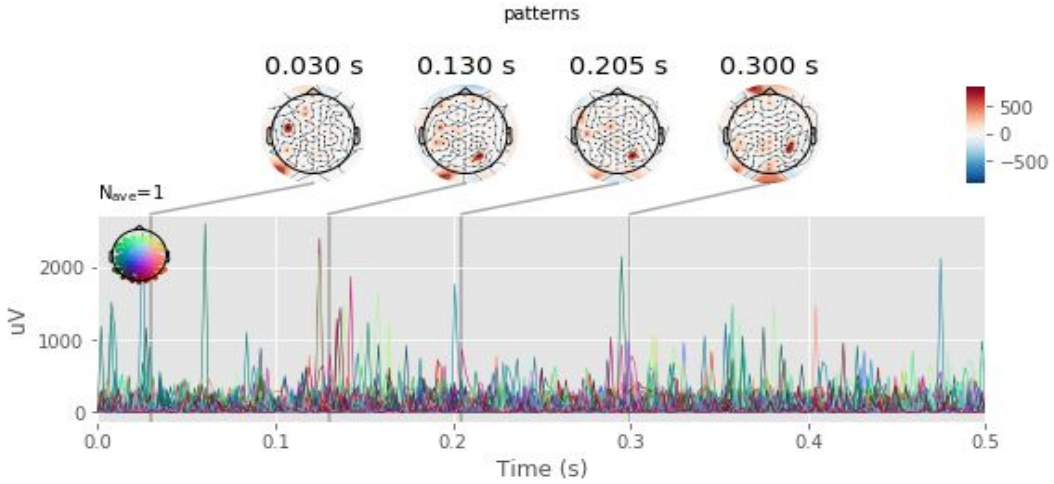


Figure 10 (top): The coefficients patterns plot over time for Language vs. Non-Language, using Logistic Regression, with topographical maps at the times with highest variance.

Figure 11 (bottom): The importance patterns plot over time for Language vs. Non-Language, using Random Forests, with topographical maps at the times with highest variance.

Conclusion

Our team was able to successfully build a classification and visualization pipeline to tackle the classification problems regarding modality (Audio vs. Visual) and lexicality (Language vs. Non-Language). Of the two binary classification models that were used, Random Forests was more accurate in its classification and had a more obvious set of channels that were most important to it during its training phase. With the method outlined in this paper, psychologists can see exactly where and when a machine learning algorithm is learning its most useful features to perform binary classification on new data. This knowledge may help psychologists better understand the decision making processes that the brain goes through in dealing with incongruent stimuli, in cases like the Stroop event.

This project, should it continue on, will need to apply the methodology outlined in this paper to better performing classification algorithms such as Linear SVM. In addition, the lexicality problem can be explored further by subsampling the data for signals that are audio only, or visual only. Furthermore, the classification problems regarding semantic context and congruence can be approached using a similar method to ours, however there may be additional preprocessing steps required. Currently, our preprocessed data removed the event IDs in the trial that corresponded to congruence and semantic content. So in the future, those IDs would have to be included in the creation of epochs from the raw data. Nevertheless, the basic structure for how the pipeline should work is in place.

REFERENCES

1. Britton, David. *Semantic Attention: Effects of Modality, Lexicality and Semantic Content.* 2017. Gradaute Center, City University of New York. New York, NY.
2. Heyden, Martin. *Classification of EEG Data using machine learning techniques*. 2016. Department of Automatic Control, Lund University. Lund, Sweden.
3. Nedelcu, Elena, et. al. *Artifact Detection in EEG Using Machine Learning.* Computer Science Department of Technical University, Cluj-Napoca, Romanian Institute of Science and Technology, CoNeural Institute Cluj-Napoca
4. King AJ, Calvert GA. *Multisensory integration: perceptual grouping by eye and ear.*
5. Laurienti P, Kraft R, Maldjian J, Burdette J, Wallace M. *Semantic congruence is a critical factor in multisensory behavioral performance.*
6. Laurienti PJ, et al. *Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices.*
7. Xingwei An, Yong Cao, Jinwen Wei, Shuang Liu, IEEE Member, Xuejun Jiao, Dong Ming*, Senior IEEE Member. *The Effect of Semantic Congruence for Visual-auditory Bimodal Stimuli*
8. Tchircoff, Andrew. The Most Complete Chart of Neural Networks, Explained. https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f236744
9. https://martinos.org/mne/stable/auto_tutorials/plot_artifacts_correction_ica.html?highlight=remove%20artifacts