Can Cross Language Brain Decoding work ?

Harshit Sharma
2019101083

Shreyash Rai
2019101096

I. Introduction

The approach of cross-language brain decoding is to use models of brain decoding from one language to decode stimuli of another language. It has the potential to provide new insights into how our brain represents multiple languages. The cross-language brain decoding approach provides a new and powerful direction to address the issue of how two or more languages are encoded through shared and distinct neural activities in the brain. These studies have significant practical implications for bilingual education and foreign language instruction. However, so far it is unclear whether and how cross-language brain decoding works, given the extant evidence. In single language brain decoding, neural responses to linguistic materials are recorded with neuroimaging methods, such as functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) and a computational model is trained to map between brain activity and stimulus-specific linguistic features.Many factors determine the accuracy of brain decoding, including the temporal and spatial resolutions in the neuroimaging recordings and the type and nature of the computational model.The approach of brain decoding of language not only helps us to understand how the brain represents language, but also has important clinical and educational implications. For example, it could be used to predict what words a person is hearing, reading or even thinking, which, in the future, could inform the design of brain-computer interfaces.An exciting new direction in recent years has been cross-language brain decoding, which is our focus here. This direction of research helps us to reveal how our brain represents multiple languages. Traditional neuroimaging studies of bilingualism have compared neural activities elicited by different languages, and identified both common and distinct neural systems of multiple languages.

II. Attempts Till Now

A number of recent studies have demonstrated that it is possible to reliably decode semantic information at the word level across different languages from neuroimaging data using machine learning methods.Multivariate pattern analysis (MVPA) has been used in cross-language decoding with increasing popularity. Compared to the traditional univariate method which examines brain voxels in isolation, MVPA takes into account the relationships across multiple voxels and has the potential to decode fine-grained patterns of brain activity. Bilinguals are usually recruited as participants in cross-language brain decoding studies and the same participants need to receive stimuli (words) from both languages (consecutively) while their brain responses are collected during the processing of these stimuli.Buchweitz, Shinkareva, Mason, Mitchell, and Just (2012) used concrete nouns from two categories (tools and dwellings) as stimuli and presented the nouns in Portuguese and English (translation equivalents) consecutively to Portuguese-English bilinguals. Participants were required to read each noun silently and think about the properties of the noun while their brain activity was recorded using fMRI. Results showed that, when the decoder was trained on the fMRI signals elicited by the English nouns and tested on the fMRI activity elicited by the Portuguese nouns, the decoding accuracy reached 0.68.

III. Factors Affecting Cross Language Decoding

As we are working on the brain signals of two different languages, we need to know how these languages are represented by our brain and how it reacts to it. Their are a number of factors to keep in mind for this.

A. *Cross Language Similarity*

Cross Language Similarity is the distance or similarity between the two languages, which may involve systematic differences in vocabulary, grammar, phonology, script, and other characteristics. Understanding the extent to which there are shared or different aspects across languages is an essential step in addressing the possible influences of language properties and linguistic experience on cross-language brain decoding.Thereby, we selected **English and Spanish** for our project. To begin with the linguistic similarities between **English and Spanish**, the most evident one is that they both share the same alphabet: the Latin alphabet. They also share a few words rooted in the old Latin and Greek languages, which greatly simplifies the comprehension of many words from the other language. For example, the words photography and fotografía, or biology and biología. These words with similar sound and meaning are called “cognates,” and comparative linguistics tells us that there are plenty of them between English and Spanish. Between 30percent and 40percent of English words have a related word in Spanish.

B. *Age of acquisition (AoA) and proficiency*

For proper results of our model, we also need to understand the dataset of whose we have the brain signals. AoA and proficiency of the second language have been found to be among the most important variables underlying the neural representation of L1 and L2 in the bilingual brain. For e.g., if we take kids in our dataset or we take people with less proficiency of l2 language, it would surely affect our study and results of the model. For our task, we took a dataset from OpenNeuro. In the dataset, L1 adult bilingual speakers (both Spanish and English) are given to read 1000 English words and their brain signals(fMRI) are recorded. Afterwards, they are given translation of these 1000 English words to Spanish and their brain(fMRI) signals are recorded.

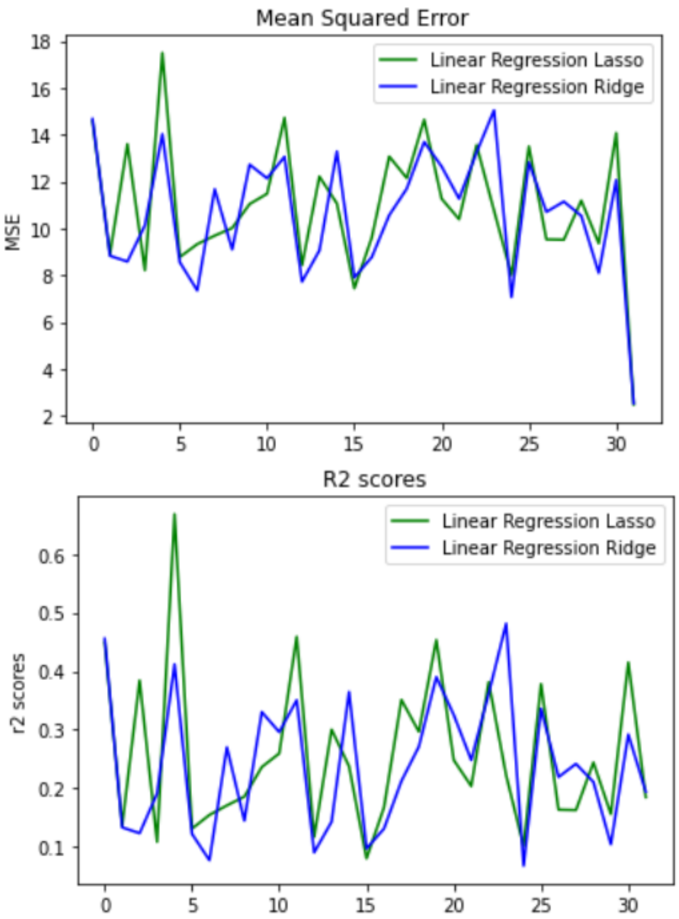
C. Depth of language processing

Brain decoding within and across languages depends on the quality of the neuroimaging data obtained, which is in turn dependent on the participant's level of processing of the language stimuli. It is well known from classic memory theories that deeper, more elaborative, and richer semantic processing would lead to better memory (e.g., more successful retrieval) than shallow or surface-level processing of the same material.Given that depth of processing can significantly impact both cross- language and within-language decoding, many previous studies, in order to engage participants in deep processing, have presented the same word stimuli for multiple times during brain imaging.For example, in Mitchell et al.'s classic study, the participants viewed the same word picture six times and their task was to think about the properties of the objects/concepts to which the stimuli refer when the stimuli were presented. The results could have been different if the stimuli were presented only one time, leading to shallow processing.

IV. Approach 1 : (Basic Approach)

In the first approach, we mapped the words to indexes using a dictionary, then using the English brain signals and indexes of english words, we trained the Linear regression model.We predicted the indexes of spanish from spanish brain signals, then calculated r2 scores and mean squared error for the same. For more better results , we used regularization and used both ridge and lasso regression for the same.The results found are :

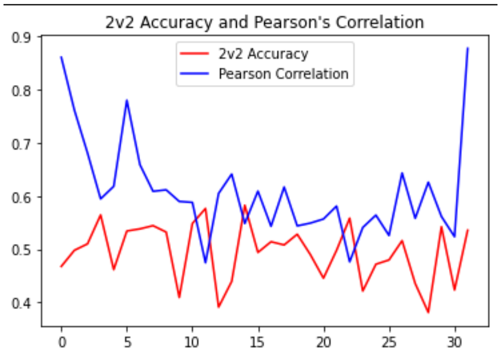
Linear Regression
R2 Score : 0.68
Mean Squared Error : 15.383
Lasso Regression
R2 Score : 0.82
Mean Squared Error : 18
Ridge Regression
R2 Score : 0.61
Mean Squared Error : 15

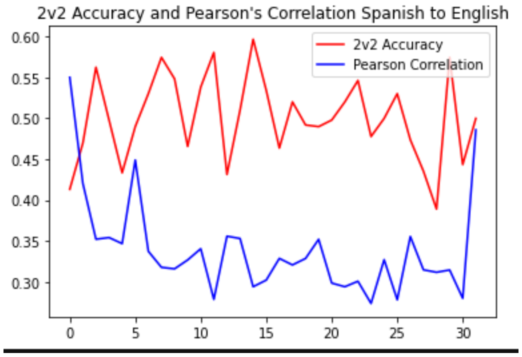


V. Approach 2 : (Using BERT encoding)

We got to know that indexes would not be a good measure to tokenize the words. So this time we wanted to do something with word embeddings. So, we used bert model of distiluse-base-multilingual-cased-v2, this is a multi-language bert embedding model. This is suitable with both English and Spanish. So we encoded our words using this bert model. After applying this, each words was mapped to 512 dimensional dense vector space. Now, we would this for our model training. We also changed our score metrics. For approach 2, we are calculating 2V2 Accuracy and Pearson Correlation, to check how good our model would perform.

A. Applying Ridge Regression





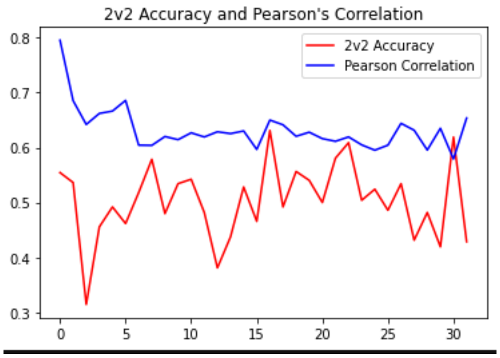
B. Random Forest Regressor

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. Therefore, we used random forest regressor with maxdepth= 2 and maxdepth=None for our model. This would give more good values than a single ridge regressor.

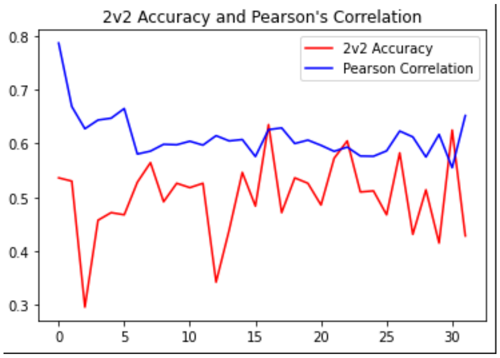
Random Forest (Maxdepth = 2 , En2Sp) : 2v2 Accuracy = 0.62 , Pearson Correlation = 0.81

Random Forest (Maxdepth = None , En2Sp) : 2v2 Accuracy = 0.63 , Pearson Correlation = 0.82

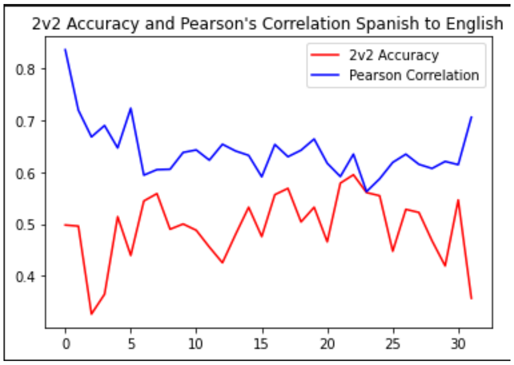
Random Forest (Maxdepth = None , Sp2En) : 2v2 Accuracy = 0.59 , Pearson Correlation = 0.84



Rf , Maxdepth = 2 , En2Sp



Rf , Maxdepth = None , En2Sp



Rf , Maxdepth = None , Sp2En

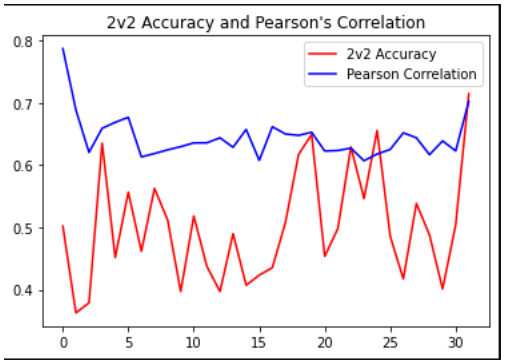
C. Bagging Regressor

A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. We need to use different regression techniques and thought that bagging would combine different prediction on multiple datasets our data.

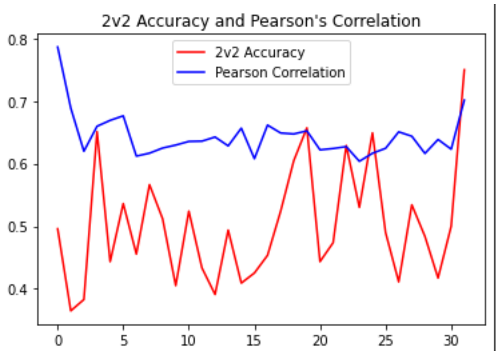
Bagging Regressor(nestimators = 50 , En2Sp) : 2V2 Accuracy= 0.71 , Pearson Correlation= 0.8

Bagging Regressor(nestimators = 100 , En2Sp) : 2V2 Accuracy= 0.75 , Pearson Correlation= 0.83

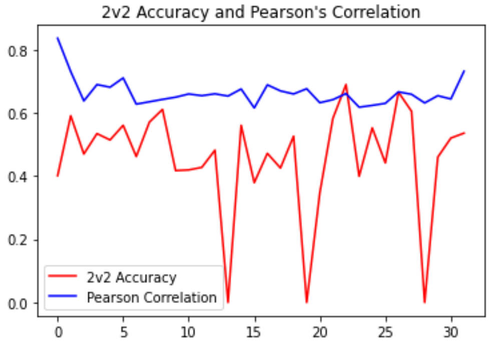
Bagging Regressor(nestimators = 100 , Sp2En) : 2V2 Accuracy= 0.63 , Pearson Correlation= 0.82



Bg, nestimators=50 , En2Sp



Bg, nestimators=100 , En2Sp



Bg, nestimators=100 , Sp2En

VI. Conclusion

Brain decoding has been an exciting and rapidly developing topic in this regard. Cross-language brain decoding has the potential to provide new insights into how our brain represents multiple languages.Our work only considered basic machine learning algorithms in the brain decoding task.We could also use advanced algos as MVPA(Multivariate Pattern Analysis) etc. for better results. Moreover, if data is available, we could also work on monolinguals rather than bilinguals, which could be an important field to explore.cross-language brain decoding indicates that it is possible to decode semantic information across different languages from neuroimaging data, but there are also significant challenges to its success. Factors such as cross-language similarity, AoA/proficiency levels, depth of language processing may all affect the effectiveness of cross-language decoding.

References

[1] Min Xu Center for Brain Disorders and Cognitive Sciences, Shenzhen University, Shenzhen 518060, China, Center for Language and Brain, Shenzhen Institute of Neuroscience, Shenzhen 518060, China

[2] Duo Liu Department of Chinese and Bilingual Studies, Faculty of Humanities, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

[3] Ping Li Department of Chinese and Bilingual Studies, Faculty of Humanities, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

[4] Edith Brignoni-Perez, Nasheed I. Jamal, and Guinevere F. Edena, An fMRI Study of English and Spanish Word Reading in Bilingual Adults