Regression techniques and Word embeddings in brain activity decoding

Dan Amler
Technion – Israel Institute of Technology
dan_amler@campus.technion.ac.il

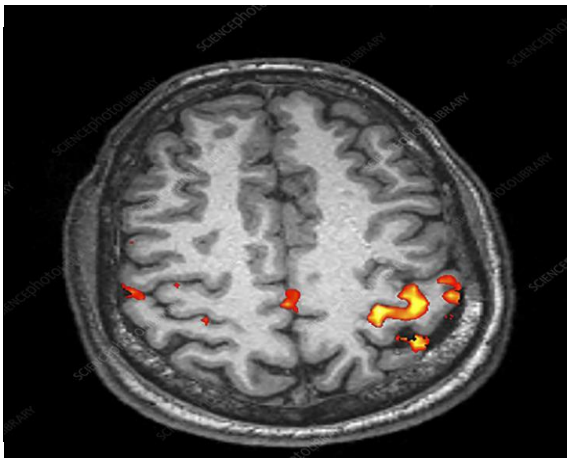
Abstract

This project investigates the decoding of semantic content from fMRI data using various word embeddings and regression models, extending the work of Pereira et al., 2018. We address the effectiveness of different regression models, including ridge, support vector regression (SVR), partial least squares (PLS), principal component regression (PCR), in predicting neural representations from linguistic data. Our analysis demonstrates that contextualized embeddings, such as those from BERT, outperform static embeddings in decoding tasks. We conclude that the choice of regression method impacts decoding accuracy, with different techniques better suited for different embeddings and tasks. The findings contribute to understanding how different machine learning approaches can optimize brain decoding processes. Github repository of used code

Introduction

Understanding how the brain processes language is one of the main challenges in cognitive neuroscience. Functional Magnetic Resonance Imaging (fMRI) is a key tool that allows researchers to observe which parts of the brain are active when a person is engaged in language related tasks. However, interpreting these complex brain signals and linking them to specific words or sentences is a difficult task.

In the field of Natural Language Processing (NLP), word embeddings are techniques used to represent words and sentences as numerical vectors. These vectors capture the meaning of words by analyzing their usage in large text corpora. Models like Word2Vec and GloVe have been developed to create these embeddings, making it possible for computers to understand and process human language in a meaningful way. Word embeddings are crucial in various NLP applications because they preserve the semantic relationships between words in a way that can be easily manipulated by machine learning algorithms. In their 2018 study, Pereira et al. explored the connection between word embeddings and brain activity. They demonstrated that it is possible to decode or predict the meaning of words from fMRI data using word embeddings. Specifically, they used GloVe embeddings to match patterns of brain activity to the meanings of words and sentences that subject were processing during the fMRI scans. Their research showed that not only individual words but also entire sentences could be decoded from brain activity using these embeddings. This was a significant step forward in understanding how the brain represents language. It suggested that the brain's processing of language could be



fMRI scan

captured by models originally developed for NLP. The study by Pereira et al. laid the foundation for further research into how different types of word embeddings might be used to better understand the relationship between brain activity and language processing. We expanded this study by testing different embeddings, including contextualized embeddings and 3 additional regression techniques to decode the fMRI scans.

Data

The data used in this project comes from the study by Pereira et al. and from pre-trained word embedding models. The data contains Functional Magnetic Resonance (fMRI) scans, that measure brain activity by detecting changes in blood flow in the brain. The fMRI data includes neural responses recorded from 3 different experiments using different concepts and sentences. This data is recorded as Voxels, 3D coordinates that captures the blood oxygen level-dependent response. Each concept or sentence is represented as a GloVe embedding. The pre-trained GloVe embedding model used in the study is GloVe 840B, trained on 840 billion tokens from the Common Crawl corpus with a vocabulary of 2.2 million words, provides 300-dimensional embedding for each word.

There were 16 participants in experiment 1. This experiment's data consists of 180 concepts, including nouns, verb, adjectives and adverbs. For each concept there is the corresponding fMRI scan that is represented as a vector of 185866 voxels. Each concept is also represented as a GloVe embedding.

There were 8 participants in experiment 2. Experiment 2 data consists of 96 text passages, each passage is related to one out of 24 topics. Each passage contains 3-4 sentences, total of 384 sentences. The brain activity fMRI scans were recorded while the participants read each sentence. Each sentence is represented by the average GloVe embedding of each word in the sentence.

There were 6 participants in experiment 3. This experiment was similar to experiment 2 except the use of 72 passages that covered 24 new topics that are unrelated to the previous 24 topics.

Additional word embeddings models were used for further analysis that was not conducted in the original study. We used the pre-trained Word2Vec model GoogleNews-vectors-negative300, which was trained on approximately 100 billion words from the Google News dataset, containing around 3 million unique words and phrases. Each word represented as a 300-dimensional vector.

We also used contextualized embeddings. PreTrained BERT embedding that was trained on the BooksCorpus that contains 800 million words and English Wikipedia (2.5 billion words), each word represented as a 768-dimensional vector. GPT2 embedding was trained on dataset of 8 million web pages (WebText) scraped from the internet. Words are also represented as 768 dimensional vector.

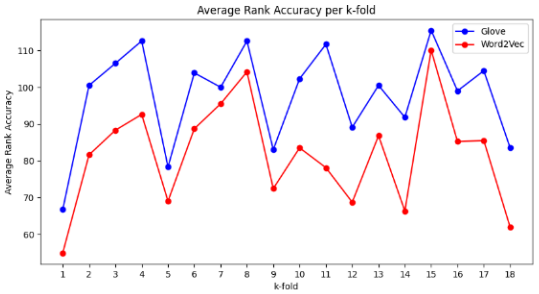
Experiments and Results

Structured Task

Sentence Decoding with Different static Embeddings:
As part of the course "Language, Computation and Cognition" we replicated a simplified version of the model from Pereira et al. (2018) to decode brain activation data. In the homework, we used GloVe embeddings to decode 180 concepts, which were words paired with pictures, from fMRI data. The goal was to predict which word the subject was looking at based on their brain activity.

We wanted to see if using a different type of word embedding, Word2Vec, would give better results compared to GloVe. To make sure the comparison was fair, we used the same dataset and applied the same method as in the homework. The data was split using an 18-fold cross-validation approach. In each part, we trained the decoder on 170 concepts and tested it on the remaining 10 concepts. This method allowed us to train and test the decoder on different parts of the data, making sure that our results were reliable.

Results:



embeddings. It's important to note that certain concepts were better decoded by using GloVe than by Word2Vec. This suggests that different types of word embeddings may capture different aspects of the semantic information

Figure 2: Average Rank Accuracy per fold for both embeddings

Testing the GloVe-Based Decoder on New Datasets:

In this part we used the GloVe-based decoder that was trained on the 180 concepts from the previous part on the datasets from analyses 2 and 3 of Pereira et al., 2018. The goal is to see how training on these 180 concepts helped the decoder to learn the underlying structure of the embedding space, enabling it to generalize the decoding for full sentences instead of concepts that are represented by 1 word. The FMRI data was structured the same way as in the previous part and the decoding process was the same.

To evaluate the performance of the decoder, for each decoded vector $v_{decoded}$, we calculated the cosine similarity to all the original embedding vectors $\{v_1, v_2, \dots, v_n\}$. For $i \in \{1,2,..,n\}$

$$\cos(v_{decoded}, v_i) = \frac{v_{decoded} \cdot v_i}{||v_{decoded}|| ||v_i||}$$

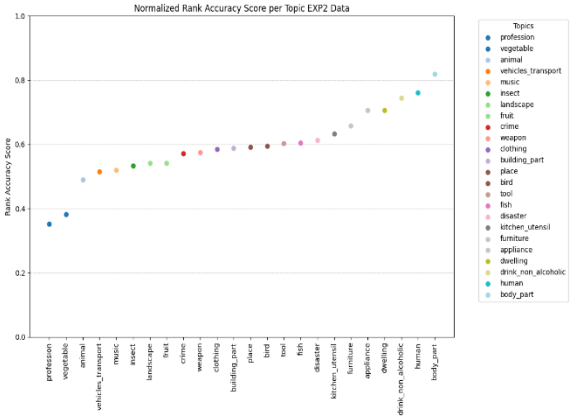
Then we ranked all the cosine similarities in a descending order, where rank 1 corresponds to the highest similarity. The score of each decoded vector is the rank of the cosine similarity to the embedding vector of the sentence. We calculated the topic-score for each topic, n_i is the number of sentences in topic i:

$$topic - score(topic) = \frac{\sum_{v_i \in topic} score(v_i)}{n_i}$$

We normalized it to get Accuracy between 0 to 1. N is the total number of embedding vectors

$$Accuracy = 1 - \frac{(topic - score) - 1}{N}$$

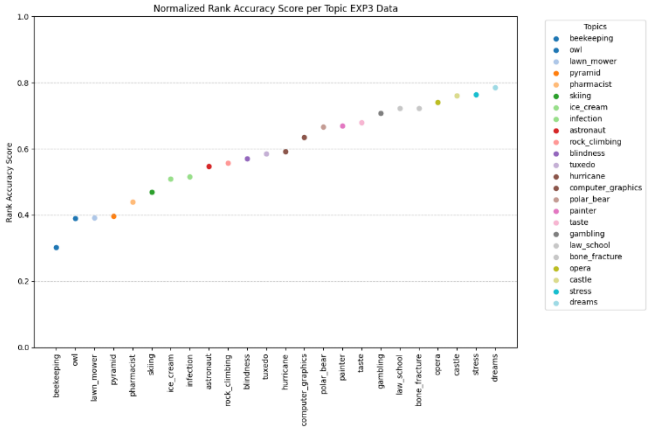
Results for experiment 2 data:



We got the highest normalized accuracy score for the topics: Body part, Human, Drink, Dwelling and Appliance. The lowest normalized accuracy score are for the topics: Profession, Vegetable, Insect.

Results for experiment 3 data:

Figure 1: Normalized Rank Accuracy for each topic (Experiment 2 data)



Here the highest normalized accuracy score is for the topics: Dreams, Stress, Castle, Opera. The worst are for: Beekeeping, Owl, Lawn Mower, Pyramid.
The difference in the accuracy score across the topics can be attributed to different reasons.

Figure 3: Normalized Rank Accuracy for each topic (Experiment 3 data)

The syntactic complexity of the sentences may vary, more complex sentences activate broader and more distributed area of the brain. (1)
Topics that involve concepts with similar neural representations may lead to lower accuracy, for example “Beekeeping” and “Insect” may activate similar brain regions, the decoder may find it challenging to differentiate between these topics.

The 180 concepts from the training data also effect the effectiveness of the decoder in predicting new topics. The decoder is better at predicting topics that are semantically or neurologically similar to the concepts that he was trained on.

Overall, some topics from experiments 2,3 were well-predicted using the concept training data, we can assume why certain topics are decoded better than others but further analysis is needed to come to conclusion.

Semi-structured Tasks

Comparing static embeddings to contextual embeddings:

Static embeddings, like GloVe, give each word a fixed vector, while contextual embeddings, like BERT or GPT-2, generate different vectors based on the word's context in a sentence, capturing more semantical meanings.

In this section, we explore the effectiveness of different sentence representations for decoding neural data. We compared the GloVe embeddings with the pre-trained contextual embeddings from BERT and GPT2, both available through the HuggingFace library. We used the dataset from experiment 2, applying 12-folds cross-validation approach for training and testing to evaluate the difference in performance of the embeddings.

Results:

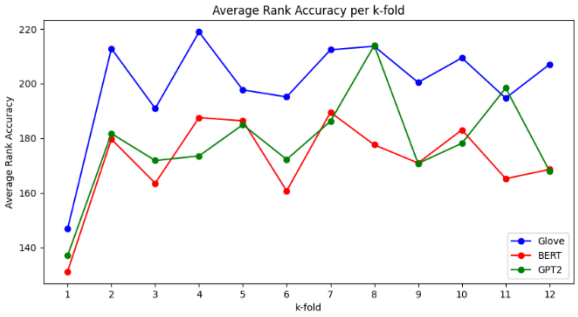


Figure 4: Average Accuracy per fold for each embedding type

This graph illustrates the effectiveness of contextual word embeddings in decoding brain activity. The average rank accuracy for BERT embeddings was the best across all 8 out of 12 folds, while GPT-2 was the best at 3 out of 12 folds, GloVe was the best for only 1 fold. GPT2 is a unidirectional model, generating embeddings based only on the preceding context in the sentence. BERT is bidirectional, considering both preceding and following context, which allow for better understanding of the meaning of the sentence. This may be the cause for BERT giving better results. (2)

Brain encoder model

In this part of our work, we switched from decoding brain activity to testing how well we can predict brain activity using word embeddings. For each voxel in the brain, we trained a regression model that used word embeddings to predict the voxel's activation. This means we tried to see how well the meaning of words, represented by word embeddings, can explain the brain's response. By training these models, we aimed to understand how the brain encodes the meaning of words at the level of individual voxels. training value. We used the log R2 score to evaluate the model.

Results:

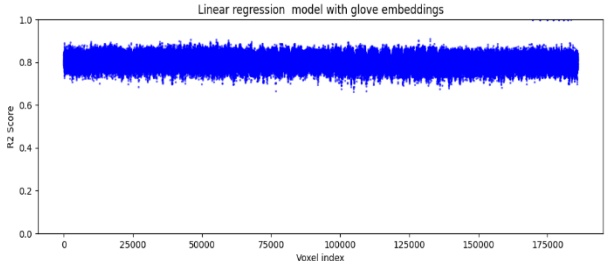


Figure 5: R2 score of linear model with Glove embedding

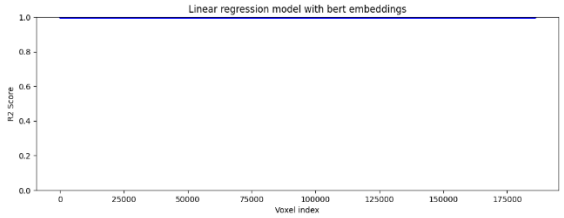


Figure 6: R2 score of linear model with BERT embedding

Using the F score statistical test on the linear regression models trained on glove embeddings, we have found a proportion of 0.1605 among the voxels to be significant (using alpha=0.05 for the p-value test). It should be noticed that some voxels were artificially easy to predict: the top voxels in the 175,000 range at the first plot have an R2 score of 1, because they were always set with the same value across the entire dataset. As we can see, the bert embeddings seem to allow for better predictive power using linear regression, but it seems highly likely this could be attributed to overfitting due to the high dimensionality of the bert embedding (700+ size vectors in BERT embeddings vs 384 samples in the Pereira dataset). Because of the BERT embedding's high dimensionality expressiveness, we chose it to go forward and activated different regressions on the training data to see if we can enhance the predictive power over the training data using BERT embeddings, while also getting more generalized model, less prone to overfitting than linear regression. While the results varied in quality, the SVR model, despite showing middling results, might be more significant when considering the danger of overfitting in a simple linear regression when the feature size is larger than the sample size Overall we can see some promising success in the task of encoding textual data to voxel form, and with bigger datasets in the future use of stronger models (namely Neural Networks) might be a promising prospect.

Open ended task

Introduction to Regression in Decoding

In the original work by Pereira et al. (2018) as in the structured and semi structured task we have completed, regression model was used to decode brain activity into semantic vectors, providing a tool for interpreting neural data. Ridge regression was employed to map high dimensional voxel data from fMRI scans to lower-dimensional semantic vectors, such as those obtained from word embeddings like GloVe, Word2Vec and BERT.

The choice of regression method affects the accuracy and reliability of the decoding process. Different regression models can handle the nuances of neural data in different ways. Some methods might be better for handling high dimensionality of voxel data while others might be better at capturing non-linear relationships between brain activity and semantic content. By selecting and tuning the right regression model we can improve the decoder's ability to accurately predict the semantic content associated with specific patterns of brain activity. In the following section we will explore how different regression methods effect brain activity decoding.

Overview of Regression Techniques

Here we will explain the applications and the differences between the different regression techniques that will be used in the following section:

Ridge Regression is a type of linear regression that adds a regularization term to the L2 norm of the coefficients. This regularization helps to prevent overfitting by shrinking the coefficients, making the model more robust to dealing with large number of features, which should help to decode the brain activity which consist of large number of voxels. (3)

Support Vector Regression (SVR) is a regression method based on Support Vector Machines (SVM) that fits a function within a margin of tolerance, using only the most important data points. This regression can handle both linear and non-linear relationships through kernels such as RBF. For decoding high dimensional brain activity data SVR can effectively model complex patterns and non-linear relation between brain activity and word embeddings. (4)

Principal Component Regression (PCR) is a regression method that combines Principal Component Analysis (PCA) with linear regression. PCR first reduces the dimension of the data by extracting the principal components, which are linear combinations of the original features that capture the most variance. It then performs linear regression using these principal components as predictors. This approach is useful for decoding high-dimensional brain activity data because it simplifies the data, reduces noise and prevents overfitting. (5)

Partial Least Squares (PLS) regression is a method that finds latent variables that capture the maximum covariance between the predictors and the responses. It projects both the input data and the output data to these latent variables, creating a lower dimensional representation that captures the most relevant information for prediction. PLS can effectively identify the latent structures that link brain activity patterns to semantic dimensions. (6)

Experiments and Results

In this section we will explain the experiments that were conducted to build on the results from the previous tasks, focusing on how different regression methods performed in decoding brain activity fMRI data to word embeddings.

Open-Task Experiment 1: We repeated the analysis from the structured task by applying k-fold cross-validation on the data from experiment 1, using static word embeddings. This was done with four different regression methods to compare their performance and to evaluate whether specific types of embeddings are better suited for particular regression techniques.
Results:

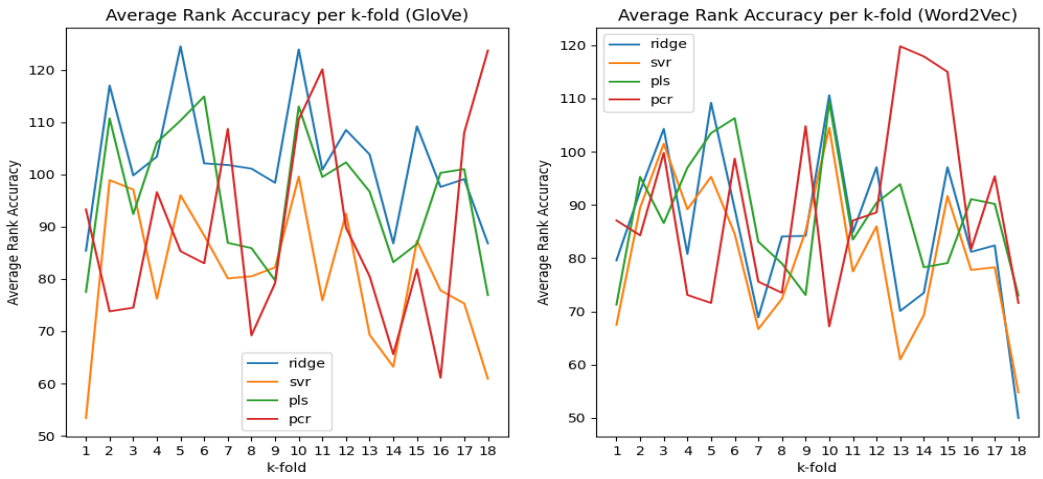


Figure 7: Average Rank Accuracy for each fold, GloVe and Word2Vec

Fold	Best Word2Vec Regression	Best GloVe Regression
1	svr	svr
2	pcr	pcr
3	pls	pcr
4	pcr	svr
5	pcr	pcr
6	svr	pcr
7	svr	svr
8	svr	pcr
9	pls	pcr
10	pcr	svr
11	svr	svr
12	svr	pcr
13	svr	svr
14	svr	svr
15	pls	pcr
16	svr	pcr
17	svr	svr
18	ridge	svr

For GloVe embeddings, the most effective regression techniques across the folds are SVR and PCR, both were optimal in 9 out of 18 folds. Ridge and PLS weren't the best in any fold.

For Word2Vec embedding, SVR was the most effective regression being optimal in 10 out of 18 folds. PCR was less effective with this embedding being optimal in only 4 out of 18 folds, PLS in 3 and ridge in 1.

Open-Task Experiment 2: We repeated the analysis from the structured task by training the decoders on the data from experiment 1, the 180 concepts, and to test it on the sentences from experiment 2. The goal is to evaluate whether specific types of regressions are better

To analyze and visualize the results, we generated plots showing normalized rank accuracy for each topic across all regression techniques for the datasets from Pereira et al. (2018), as seen in Figures X and Y. We also created tables for each regression method, highlighting the topics with the highest and lowest decoding accuracies to identify which topics were best and worst decoded.

Some topics are consistently well decoded across all regression methods. For instance, *Human* shows consistently high accuracy with all decoders, indicating that this topic's characteristics are effectively captured by each method. Conversely, some topics are poorly decoded across most decoders; for example, *Lawn Mower* performs poorly with Ridge, SVR, and PLS. Additionally, it is notable that certain topics are decoded effectively by some decoders but poorly by others. For instance, *Profession* ranks among the top four topics for PCR but is among the worst for other decoders. This variability highlights that the effectiveness of different regression models can vary significantly depending on the topic of the fMRI data.

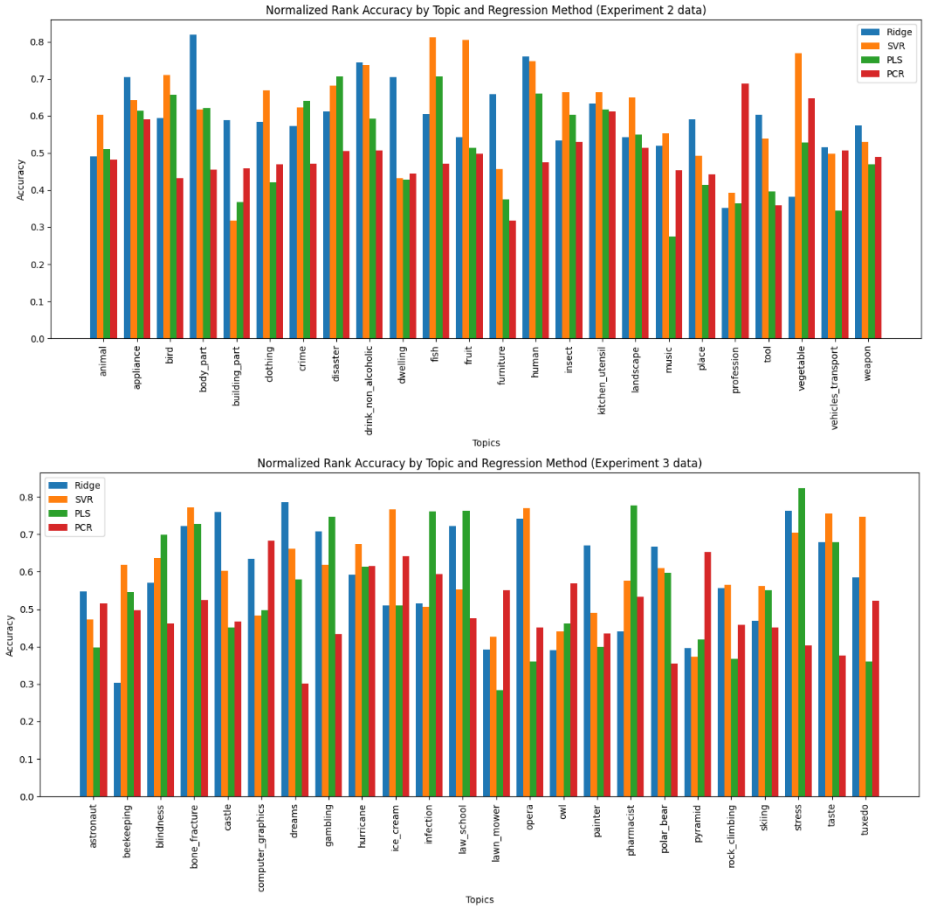


Figure 8,9: Normalized Rank Accuracy by Topic, datasets 2 and 3

Experiment 2 data:

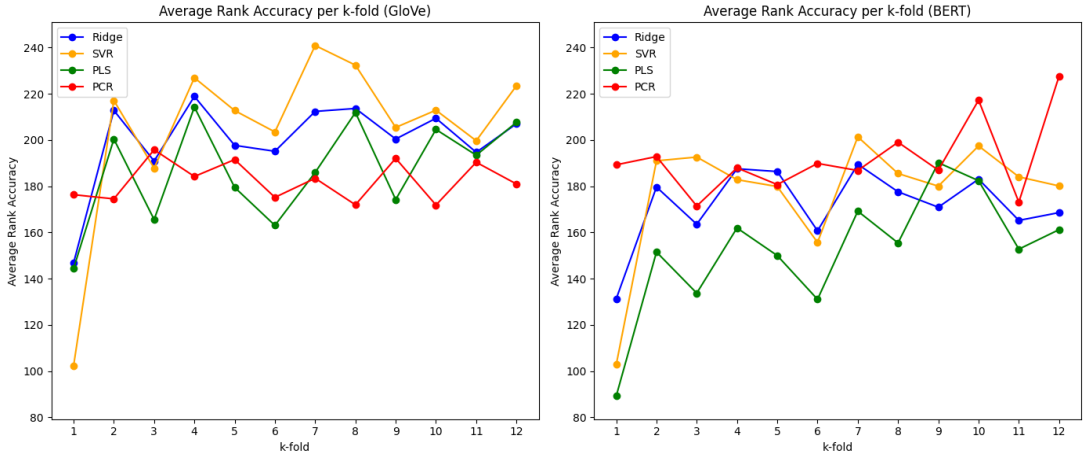
Regression type	best accuracy	Worst accuracy
Ridge	Body_part Human drink_non_alcoholic dwelling	Profession Vegetable Animal Vehicles_transport
SVR	Fish Fruit Vegetable Human	Bulding_part Profession Dwelling Furniture
PLS	Fish Disaster Human bird	Music Vehicles_transport Profession Building_part
PCR	Profession Vegetable Kitchen_utensil appliance	Furniture Tool Bird place

Experiment 3 data:

Regression type	best accuracy	Worst accuracy
Ridge	Dreams Stress Castle Opera	beekeeping owl lawn_mower pyramid
SVR	Bone_fracture Opera Ice_cream taste	Pyramid Lawn_mower Owl Astronaut
PLS	Stress Pharmacist Law_school infection	Lawn_mower Opera Tuxedo Rock_climbing
PCR	Computer_graphics Pyramid Ice_cream hurricane	Dreams Polar_bear Taste stress

Open-Task Experiment 3: We repeated the analysis from the semi-structured task of k-fold cross-validation training on the data of experiment 2 using the contextualized word embeddings BERT. The goal is to evaluate how different types of regression techniques effect the decoding on contextualized word embeddings and to see how their effectiveness compares to the static GloVe embedding. Because of long computation time of the SVR regression on BERT (2 hours for 1 fold) GPT-2 embedding was not tested.

Results:



Fold	Best BERT Regression	Best GloVe Regression
1	pls	svr
2	pls	pcr
3	pls	pls
4	pls	pcr
5	pls	pls
6	pls	pls
7	pls	pcr
8	pls	pcr
9	ridge	pls
10	pls	pcr
11	pls	pcr
12	pls	pcr

PLS consistently outperforms other models for BERT across the majority of folds, indicating that PLS regression is particularly well-suited for use with BERT embeddings. While PLS shows moderate success with GloVe embeddings, the performance for GloVe is more varied, with PCR often leading, but PLS and SVR also demonstrating strength in certain folds. This variability suggests that GloVe embeddings are more sensitive to the choice of regression model, possibly due to differences in noise characteristics or data distributions across folds. These findings, along with previous observations in this section, underscore the importance of selecting the appropriate regression model based on the specific embedding and data characteristics.

Discussion and Conclusions:

The findings in this project provide insights into how different word embeddings and regression techniques affect the decoding of brain activity from fMRI data.

Impact of Word Embeddings:
The results demonstrate that contextualized embeddings, particularly BERT, outperform static embeddings like GloVe in decoding brain activity. This superiority is likely due to the ability of contextual embeddings to capture semantic information by considering the surrounding context of words within a sentence. BERT, being bidirectional, can leverage both preceding and following contexts, resulting in more accurate representation of the semantic content. These findings supports the idea that human language processing is context dependent.

Influence of Regression Techniques:
The comparison of different regression models demonstrates how the effectiveness of the technique in decoding brain activity is highly connected to the data and the word embedding which is chosen. The regression models used in this work were chosen for their effectiveness in complex and high dimensional data with non linear relationships
Ridge, the original regression technique used in Pereira et al. (2018) performed well on open-task-experiment-2, on the other experiments it was less effective than other techniques.
SVR was supposed to handle non-linear relationships between fMRI scans and word embeddings, also being effective in high dimension. It was highly effective with Word2Vec embedding and somewhat effective with GloVe, but it was definitely not the best choice for contextualized embeddings.
PCR was supposed to focus on the most informative components of the data while filtering the noise which is important for high dimensional data and noise of fMRI scans. It was somewhat effective on GloVe embedding but performed relatively poorly on open-experiment-2.
PLS is effective at capturing complex latent relationships, it was the best choice for the BERT embedding, performed well on Word2Vec and on open-experiment-2.

We demonstrated that different regression models bring distinct strengths to various machine learning tasks, making them more or less suited depending on the nature of the data and the word embeddings used. This implicates the need to experiment with different regression models in machine learning tasks, especially in NLP tasks that utilize word embeddings. Understanding the functionalities and strengths of a regression model alone does not suffice to predict its performance for a particular task. The clear limitation of this work is that we do not fully understand why the performance of a specific regression technique changes with different word embeddings or tasks. Although we have observed that various models have different strengths, the reasons behind these performance differences are not yet clear. Future work should focus on exploring why these variations occur. This could involve looking more closely at how the features of word embeddings and tasks affect the performance of different regression models. Additional research could help us understand these interactions better.

Refrences:

Pereira, Francisco, et al. "Toward a universal decoder of linguistic meaning from brain activation." *Nature communications* 9.1 (2018): 963.

(1) Pallier, Christophe, Anne-Dominique Devauchelle, and Stanislas Dehaene. "Cortical representation of the constituent structure of sentences." *Proceedings of the National Academy of Sciences* 108.6 (2011): 2522-2527.

(2) Devlin, Jacob. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

(3) Marquardt, Donald W., and Ronald D. Snee. "Ridge regression in practice." *The American Statistician* 29.1 (1975): 3-20.

(4) Cortes, C. "Support-Vector Networks." *Machine Learning* (1995).

(5) Massy, William F. "Principal components regression in exploratory statistical research." *Journal of the American Statistical Association* 60.309 (1965): 234-256.

(5) Geladi, Paul, and Bruce R. Kowalski. "Partial least-squares regression: a tutorial." *Analytica chimica acta* 185 (1986): 1-17.