CSAI

Assignment-5

Report

Akshit Sharma
(2021101029)

Interpretation of Performance metrics

2V2Acc

It is a metric used to assess the performance of encoder/decoder models by measuring how well they capture the nearest neighbour relationships between the outputs (word embeddings or fMRI data). It essentially calculates the proportion of times the nearest neighbour of a word vector/voxel data in one model is also the nearest neighbour of the corresponding word vector/voxel data in another model, and vice versa. 2V2Acc ranges from 0 to 1, where 0 indicates no similarity and 1 indicates perfect similarity.

While 2V2Acc is typically used to compare two separate models (Model A and Model B), the concept can be extended to assess the performance of a single model by comparing its predicted embeddings (y^) to the true embeddings (y), which might be available in certain evaluation settings. Given a word vector (y^), it first performs a nearest neighbour search in its own embedding space to find the word vectors most similar to y^ using a distance metric (cosine similarity). It then compares these nearest neighbours to the true embedding (y) for the same word. The proportion of times the nearest neighbour of the predicted embedding (y^) is also the nearest neighbour of the true embedding (y) is what the metric represents.

Pearson Correlation

Pearson correlation, also known as the Pearson correlation coefficient (r), is a statistical metric used to measure the strength and direction of the **linear relationship** between two continuous variables. It provides a value between -1 and +1, where:

- **-1 indicates a perfect negative correlation:** As the value of one variable increases, the value of the other variable decreases perfectly in a linear fashion. Imagine a straight line with a negative slope.
- **0 indicates no correlation:** There is no linear relationship between the two variables. The data points would be scattered randomly with no discernible pattern.

- **+1 indicates a perfect positive correlation:** As the value of one variable increases, the value of the other variable increases perfectly in a linear fashion. Imagine a straight line with a positive slope.

Important points to consider when interpreting Pearson correlation:

- **Linearity:** Pearson correlation only measures linear relationships. It doesn't capture non-linear relationships, even if they exist between the variables.
- **Direction:** The sign of the correlation coefficient tells you the direction of the relationship (positive or negative).
- **Magnitude:** The absolute value of the coefficient (without the sign) indicates the strength of the relationship. However, it doesn't tell you how much one variable changes in response to a change in the other. Correlation does not imply causation: Just because two variables are correlated doesn't necessarily mean that one causes the other. There could be a third underlying factor influencing both variables.

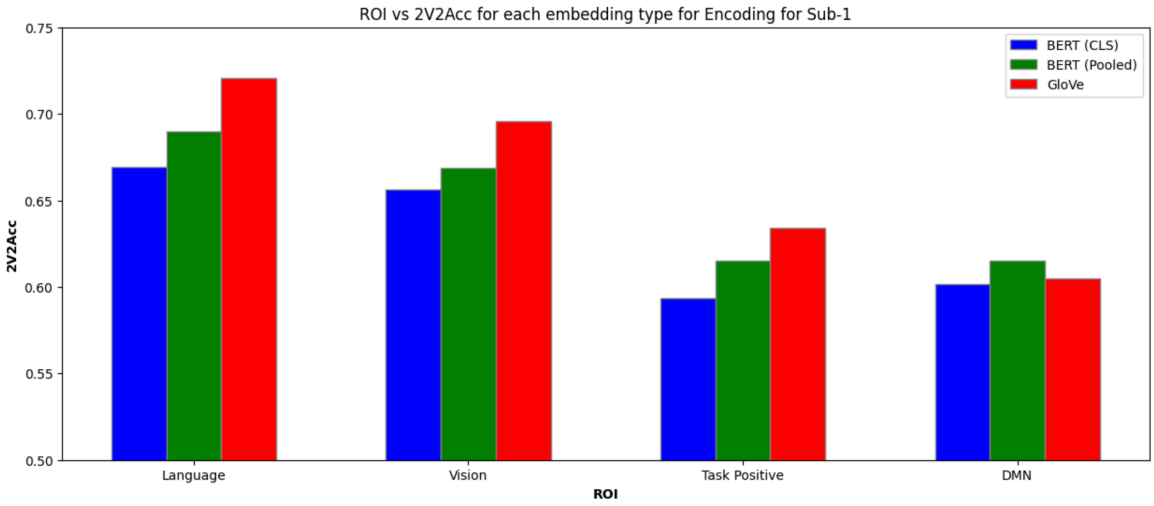
So, based on the above details of the two metrics, we can say that:

- PC measures the linear relationship between two variables. In the context of fMRI data, Pearson correlation can be used to assess the similarity of brain activation patterns between different regions or between different experimental conditions. It quantifies how much two brain regions or conditions covary in their activation levels over time. So, **PC is more the more useful metric to analyse performance of the encoder model (predicts fMRI data from sentence representations) than 2V2Acc**(measures only directional difference, not magnitude).
- Word embeddings often represent words in high-dimensional vector spaces. Cosine similarity is effective in high-dimensional spaces because it focuses on the angle between vectors rather than their magnitudes. This means it can capture similarity even in sparse, high-dimensional spaces where Euclidean distance might not be as informative. So, **for comparing sentence embeddings in the decoder models, 2V2Acc might be more useful than Pearson correlation**.

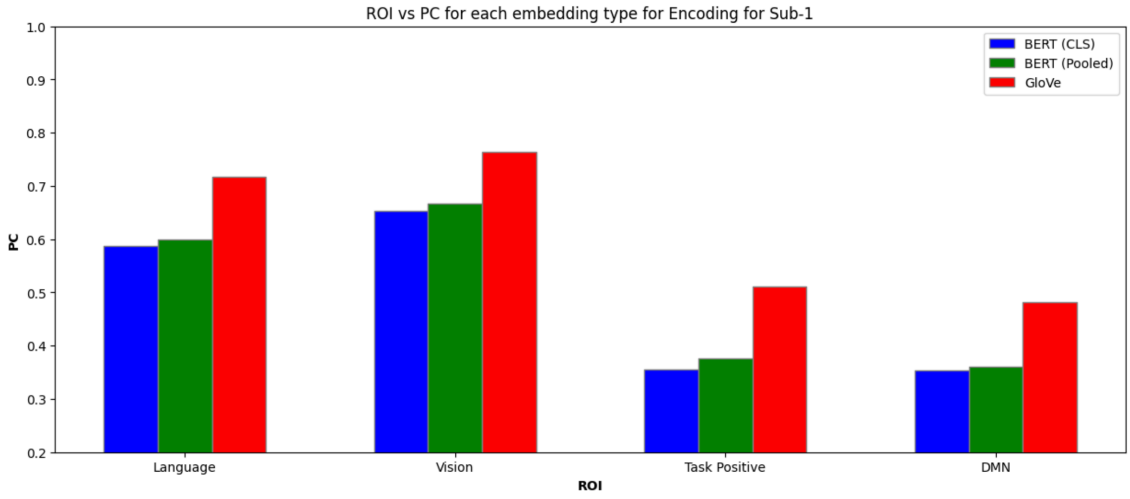
Bar Plots for 2V2Acc and PC for each ROI for 3 types of Embeddings

Subject-1

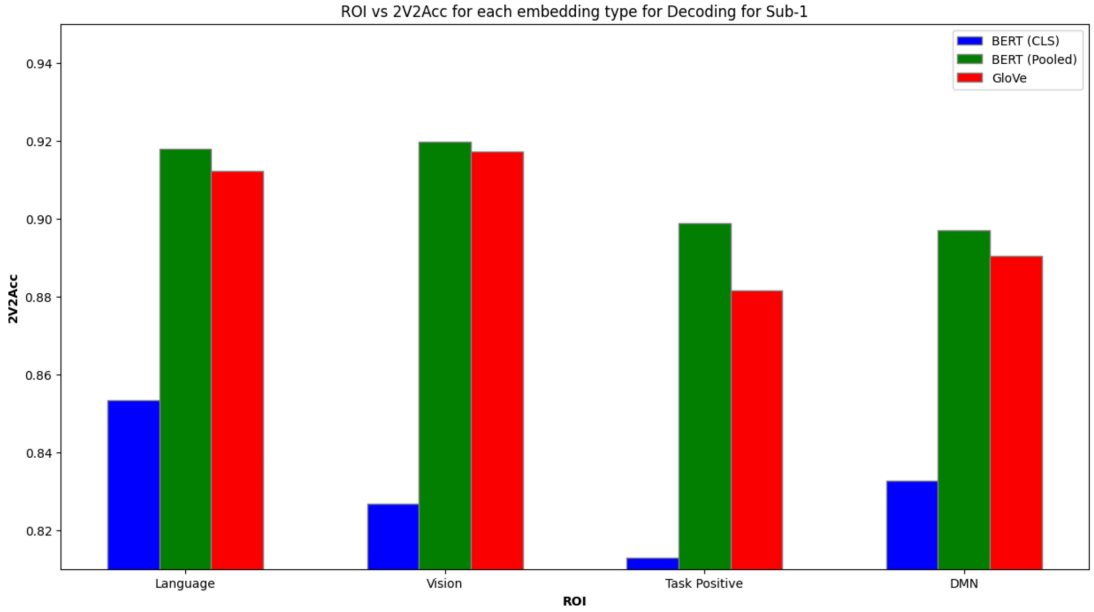
Bar plot for ROI vs mean 2V2Acc across folds for 3 types of embeddings for subject-1 for Encoding:



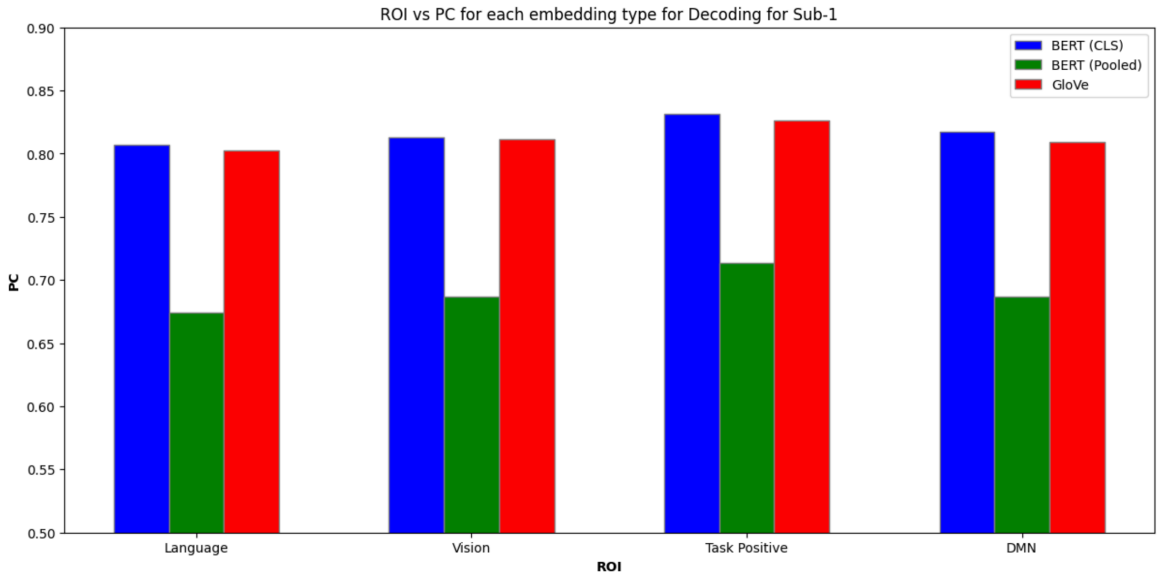
Bar plot for ROI vs mean PC across folds for 3 types of embeddings for subject-1 for Encoding:



Bar plot for ROI vs mean 2V2Acc across folds for 3 types of embeddings for subject-1 for Decoding:

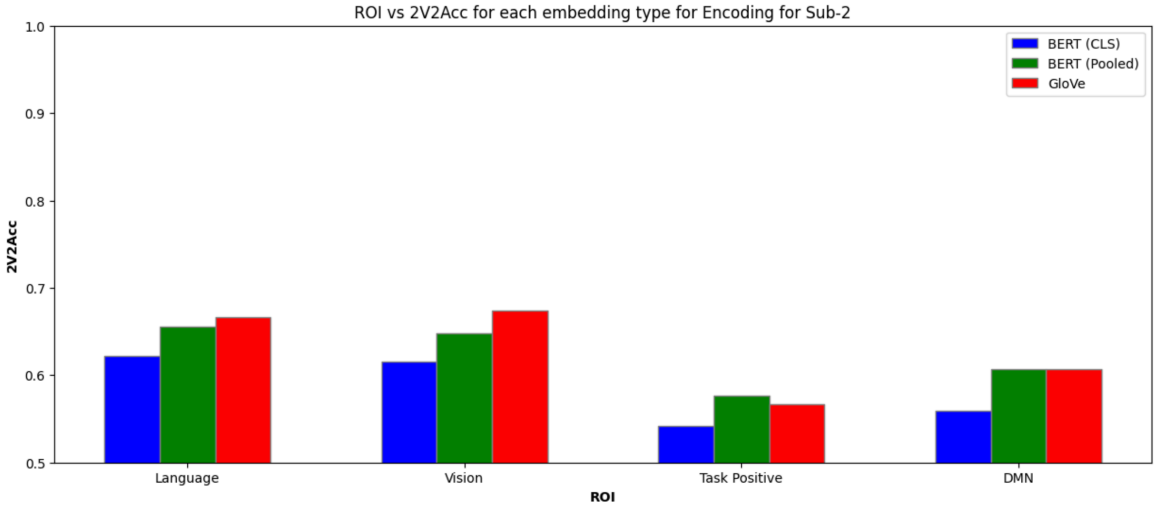


Bar plot for ROI vs mean PC across folds for 3 types of embeddings for subject-1 for Decoding:

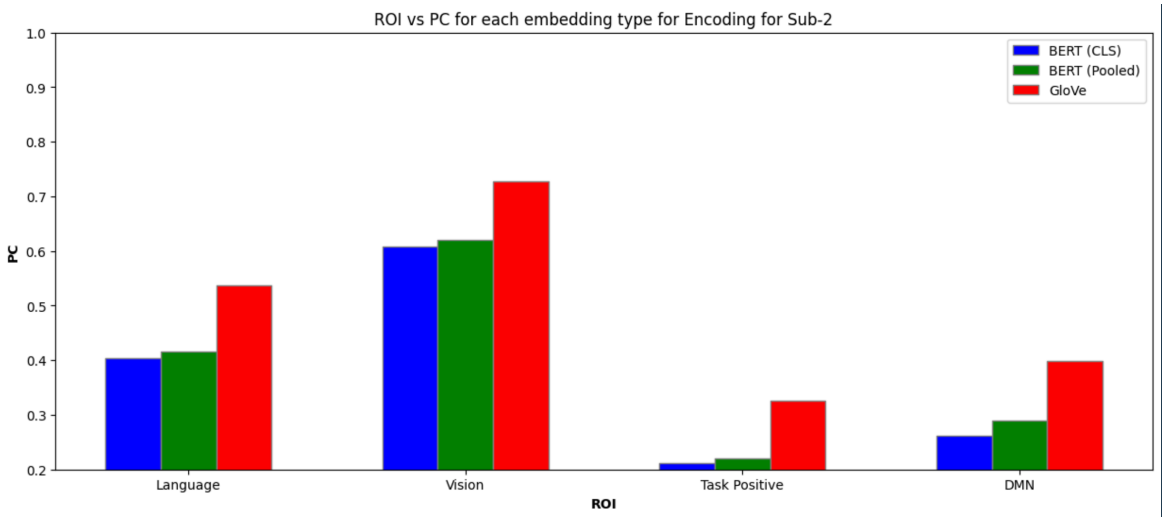


Subject-2

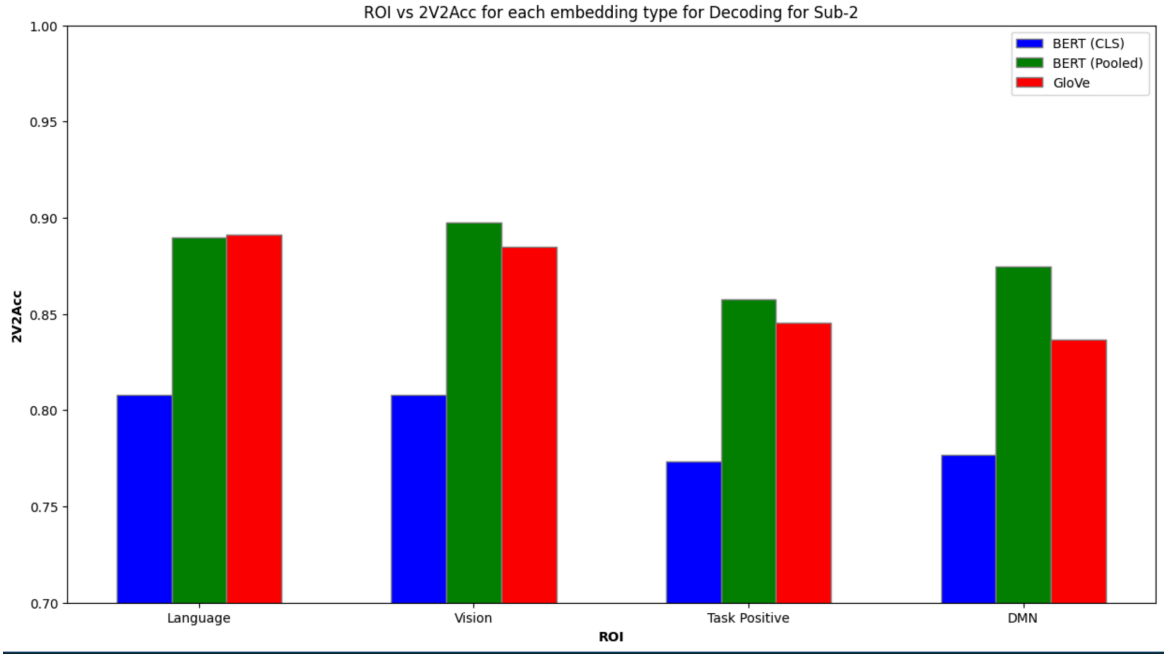
Bar plot for ROI vs mean 2V2Acc across folds for 3 types of embeddings for subject-2 for Encoding:



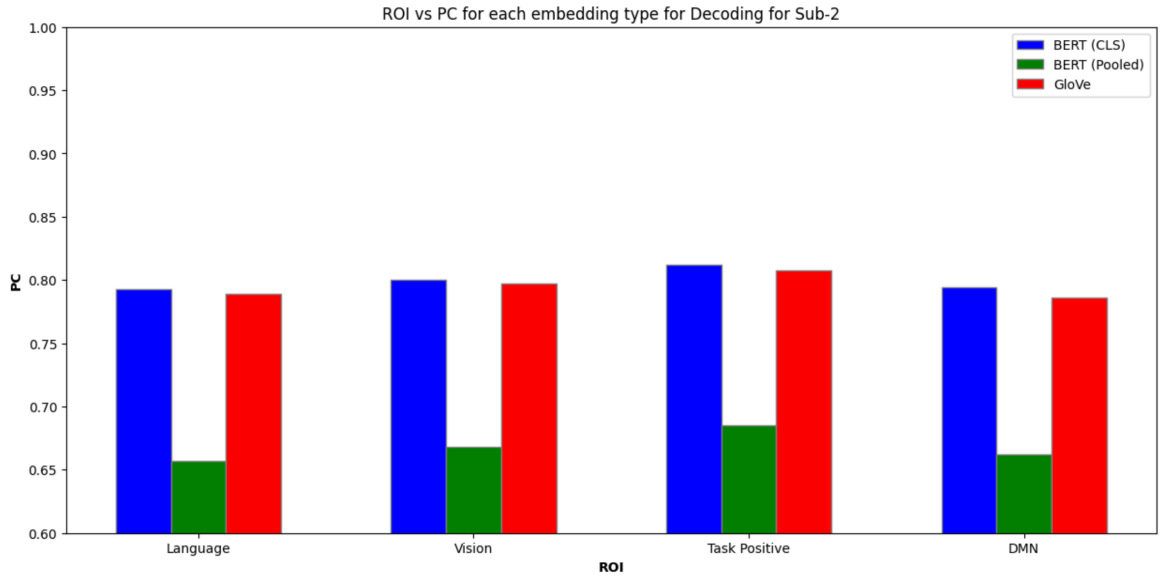
Bar plot for ROI vs mean PC across folds for 3 types of embeddings for subject-2 for Encoding:



Bar plot for ROI vs mean 2V2Acc across folds for 3 types of embeddings for subject-2 for Decoding:



Bar plot for ROI vs mean PC across folds for 3 types of embeddings for subject-2 for Decoding:



Analysis of Results obtained

Encoding

We observe that in both subjects, both the PC and 2V2Acc graphs show **high values of mean metrics in the Language and Vision ROIs**, compared to significantly lower values for DMN and Task Positive ROIs. This is because When participants are engaged in reading tasks, especially those involving comprehension or semantic processing, **language-related brain regions like Broca's area and Wernicke's area are typically activated**. Similarly, visual processing regions like the occipital cortex are engaged. The data we have here is for the task of reading text line by line, which aligns with the language and vision related brain region tasks. The default mode network (DMN) is associated with **introspection, self-referential thoughts, and mind wandering**. It is typically active during rest and less engaged during externally focused tasks. This is the reason for the low values of metrics in DMN ROI.

Also, we observe that **Glove embeddings give better performance**, followed by BERT (pooled) and the least for BERT (CLS). The "CLS" token embedding, which stands for "classification," is the representation of the special [CLS] token that BERT adds at the beginning of each input sequence. This token is used for classification tasks. However, it's a **fixed representation of the entire input sequence and might not capture all the nuances** of the sentence, especially in longer texts. On the other hand, averaging the embeddings of all the tokens in the sentence provides a **more comprehensive representation**. Each token's embedding **captures its contextual information within the sentence**, and averaging them allows the model to consider the **contributions of all words equally.**

Glove embeddings typically have a **lower dimensionality(300)** compared to BERT embeddings **(768)**. While BERT embeddings are contextualised and capture rich semantic information, they come with a high dimensionality due to the architecture of the model. In tasks where the data might **not be large or complex enough to fully utilise the contextual** information provided by BERT embeddings, the lower dimensionality of Glove embeddings might suffice and **even outperform BERT embeddings**. Glove embeddings, **being simpler and trained on co-occurrence statistics of words, might generalise better to some tasks**. This might be the case here in our encoding models, which use the **linear Ridge model** which might not be able to capture the high dimensional and complex BERT embeddings.

Decoding

We observe that in decoding too, we observe that the **Language and vision ROIs give higher values of metrics compared to DMN and Task Positive ROIs.** The reasons for this observation are similar to the ones stated above. Vision and Language ROIs are more directly

related to processing sensory input (visual and linguistic), which might **align better with the nature of the sentence embeddings.** The representations encoded in these regions might be **more discriminative and easier to decode based on the fMRI responses**. Vision and Language ROIs might show **more consistent and distinguishable patterns in response to different sentence embeddings**, making decoding more reliable. In contrast, DMN and task-positive ROIs might **exhibit more variability**, making it **harder to decode sentence representations** accurately.

As discussed in the beginning, the better metric to analyse decoding performance is 2V2Acc. We observe that the performance on BERT (pooled) is best, followed by Glove and the least for BERT (CLS). BERT embeddings are contextual, meaning they are generated based on the entire context of the sentence in which a word appears. This contextual information can capture nuances and subtle meanings that static embeddings like GloVe might miss. In the ma **encoding task, this contextual information might be crucial for accurately mapping fMRI data to sentence embeddings**. However, **in the decoding task, the static nature of GloVe embeddings might be advantageous because they provide a fixed representation of the sentence that can be more easily decoded from fMRI data without needing to consider context.** This is the reason for BERT (pooled) performing better than Glove embeddings.

Also, BERT embeddings are trained on large-scale datasets and are designed to be **robust to noise and variations in language usage**. This robustness might make them more **resilient to noise in the fMRI data**, leading to better performance in the encoding task. However, in the decoding task, where the focus is on **reconstructing the original sentence from noisy fMRI signals**, simpler embeddings like GloVe might **perform better due to their reduced susceptibility to noise**. This is the reason for **Glove performing better than BERT (CLS).**