Language, Computation and Cognition - Final Project
Lecturer: Dr. Yevgeni Berzak

Ariel Cohen and **Dan Israeli**
Technion - Israel Institute of Technology

Structured Task

Section 1

In the paper "Pereira et al., 2018", the researchers created a decoder which decodes fMRI data to embedded word vectors from the GloVe embedding model. In Homework Assignment 3 question 3, we evaluated that decoder using cross-validation (CV) on the data from experiment 1. In the CV process, we used 18 folds – each containing 10 concepts (since experiment 1's data contains 180 concepts). In this section, we were asked to repeat the evaluation process using a different static word embedding model. We chose to use the Word2Vec model, and particularly the "word2vec-google-news-300" model from the Gensim (Python) library. This model was pre-trained on approximately 100 billion words, and embeds roughly 3 million words and phrases into a 300-dimensional vector space. In the process of embedding the concepts using the Word2Vec model, we noticed that the concept "argumentatively" is not included in it. However, the word "argumentative" is included. Since the two words are very similar, we considered using the embedding of the word "argumentative" instead. After consulting with a course staff member, we were indeed advised to do so. Note that since the data contains 180 concepts, the accuracy score range is 1 (best) - 180 (worst). Moreover, this means that the expected accuracy score of a random classifier is 90.

Results

The CV evaluation results were as follows: For the GloVe model, we received a mean accuracy score of 61.91 and a median accuracy score of 65.6. In addition, the best fold accuracy score was 36.8, which was received on fold 16. Moreover, the worst fold accuracy score was 105.1, which was received on fold 9 (for the full results distribution, see Fig1). For the Word2Vec model, we received a mean accuracy score of 60.58 and a median accuracy score of 60.5. In addition, the best fold accuracy score was 36.5, which was received on fold 15. Moreover, the worst fold accuracy score was 84.0, which was received on fold 3 (for the full results distribution, see Fig2).

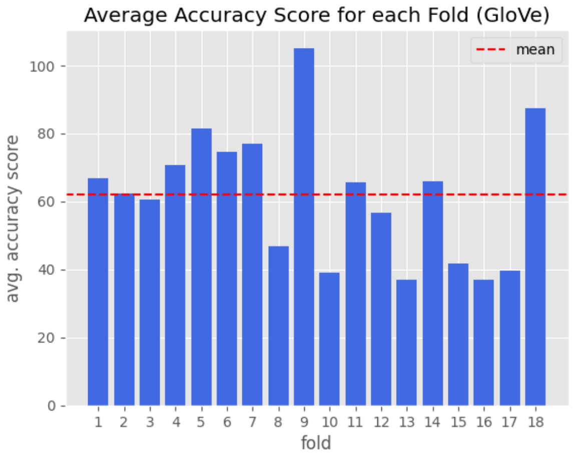

Fig 1

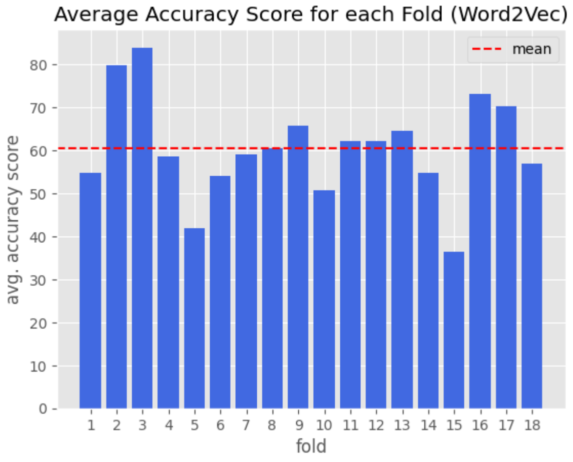

Fig 2

Discussing the Results

As we can see, the mean and median accuracy scores of both models are much better (lower) than the expected accuracy score of a random classifier (90). Thus, we can conclude that both possess a significant amount of predictive power. However, from the comparison above, it appears that the Word2Vec embedding model performed better in every measure mentioned. Therefore, the model's overall performance is better when using this embedding model. Now, let us compare the concepts that each model can predict with more/less success. We will define that our model is considered to decode a concept "with more success" if its accuracy score on the concept is strictly less than 90 (strictly better than a random classifier). We consider the accuracy score 90 on a concept to be "with less success" since a simple random classifier (that requires no training at all) can perform just as good. When using the GloVe model, the decoder successfully decodes 134 concepts (74.4%). When using the Word2Vec model, the decoder successfully decodes 135 concepts (75%). Thus, the Word2Vec model is (slightly) better than the GloVe model according to this measure as well. Moreover, we noticed that the top 5 most successfully decoded concepts are very different between the two models. The same insight goes for the top 5 least successfully decoded concepts as well. These findings further emphasize the influence of the word embedding model on the decoder's behavior.

Section 2

In the paper “Pereira et al., 2018”, the researchers conducted 3 different experiments. In this section, we will discuss the similarities and differences between them.

Similarities

1. All experiments collected fMRI data received by exposing the subjects to stimuli (concepts or sentences).

2. All experiments evaluate the performance of the suggested decoder which predicts what concept/sentence stimulus the subject was exposed to, given their fMRI data.

3. All experiments used the GloVe word embedding model.

4. In all experiments, all participants are fluent English speakers.

5. The number of topics that were used in both experiments 2 and 3 is identical (24 topics in each). Moreover, in each experiment the topics are broad.

Differences

1. The number of subjects in experiments 1, 2 and 3 is 16, 8 and 6 respectively.

2. In experiment 1, the stimuli presented to the subjects are concepts (single target words). In experiments 2 and 3, the stimuli presented to the subjects are sentences.

3. In experiment 1, each concept was presented using 3 paradigms (which were shown 4-6 times each):

   - "The target word was presented in the context of a sentence that made the relevant meaning salient".
   - "The target word was presented with a picture that depicted some aspect(s) of the relevant meaning".
   - "The target word was presented in a cloud, surrounded by five representative words from the cluster".

   However, in both experiments 2 and 3, for each passage, the sentences which are associated with that passage were shown to the subject one by one, and a total of 3 times. Moreover, each sentence was presented to the subject for the duration of 4 seconds (each time shown).

4. In all experiments the model was trained on fMRI data received from a concept stimulation. However, in experiment 1, the trained model tries to predict concepts, while in experiments 2 and 3 the trained model tries to predict sentences.

5. In experiments 2 and 3, the decoder model was trained on all of the collected data from experiment 1. However, in experiment 1, this is not the case since cross-validation was performed (so the entire data was never used all at once as a training set).

6. Experiment 2 used 96 passages, each consisting of 4 sentences. On the other hand, experiment 3 used 72 passages, each consisting of 3-4 sentences.

7. The topics that were used in experiments 2 and 3 are different. Moreover, in experiment 2 each topic was expressed by 4 passages, but only by 3 passages in experiment 3.

Section 3

In this section, we were asked to train our decoder on the data from experiment 1, and to evaluate its performance on the data from experiments 2 and 3. Notice that in experiment 1, the fMRI data is received from concept (single target word) stimulation. However, the fMRI data in experiments 2 and 3 is received from sentence stimulation. For each of the two experiments, we will find the mean and median accuracy scores of our model, as well as the number of sentences that were decoded with "more success" (according to the same principle discussed in the first section).

Results

For Experiment 2: First, note that since the data contains 384 sentences, the accuracy score range is 1 (best) - 384 (worst). Moreover, this means that the expected accuracy score of a random classifier is now 192. We found that the mean and median accuracy scores of our trained model (on all sentences) are 156.93, 136 respectively. Notice that these scores are much better (lower) than the ones of a random classifier (192). In addition, since the number of sentences (in the experiment) is 384, our threshold for a successful decoding is 192 (non-inclusive). According to this labeling criteria, 244 sentences (63.5%) were decoded with more success.

For Experiment 3: First, note that since the data contains 243 sentences, the accuracy score range is 1 (best) - 243 (worst). Moreover, this means that the expected accuracy score of a random classifier is now 121.5. We found that the mean and median accuracy scores of our trained model (on all sentences) are 100.75, 87 respectively. Notice that these scores are much better (lower) than the ones of a random classifier (121.5). In addition, Since the number of sentences (in this experiment) is 243, our threshold for a successful decoding is 121.5 (non-inclusive). According to this labeling criteria, 203 sentences (83.5%) were decoded with more success.

Discussing the Results

As we can see, in both experiments, our trained model performed much better than a random classifier. Moreover, most of the sentences were decoded successfully with our model. Thus, our model possesses a significant amount of predictive power as attested by the mentioned metrics. This insight is very impressive given the fact that our model was trained on concepts and not sentences. For that same reason, it also implies excellent generalization capabilities.

Section 4

In this section, we were asked to identify the topics where the decoder was more/less successful in predicting the sentences. According to the paper, in each of experiments 2 and 3, there are 24 different topics. Moreover, these topics are different between the two. In both experiments, each sentence is associated with a passage and each passage is associated with a topic. Thus, we will define that a sentence is associated with a topic if its passage is associated with the topic. Now, we will define that a topic is "successful" if the average accuracy score of all the sentences associated with it is better (strictly lower) than a random classifier (the same principle discussed in the first section). For each of the two experiments, we will find the number of successful topics.

Results

For Experiment 2: since the number of sentences is 384, our threshold for success is 192 (non-inclusive). According to this labeling criteria, 21 topics (87.5%) were successful.

For Experiment 3: since the number of sentences is 243, our threshold for success is 121.5 (non-inclusive). According to this labeling criteria, 18 topics (75%) were successful.

Discussing the Results

From the results above, in both experiments, most of the topics are considered successful using our model. This implies that even though our model was initially trained on concepts, it has a significant amount of predictive power on sentences from a broad range of topics. This insight is remarkable and furthermore supports our conclusion in the previous section regarding the model's excellent generalization capabilities.

Semi-Structured Task

Section 1

In this section, we were asked to train a decoder using both the GloVe word embedding model (used in the paper) and a contextualized word embedding model, in order to compare their performance. Moreover, we were asked to train the decoder using the dataset from either experiment 2 or 3. Note that experiment 2 contains more sentences than experiment 3 (384 compared to 243). Since using a larger dataset allows us to better understand the model's capabilities and performance, we chose to use this data. Moreover, we chose the BERT contextualized word embedding model from the Sentence-Transformers (Python) library. In order to assess the performance of each of the described models, we performed CV. This assessment method will allow us to gain an overall picture of our model's performance, and thus project objective findings. In the CV process, we used 16 folds - each containing 24 sentences. Note that since the data contains 384 sentences, the accuracy score range is 1 (best) - 384 (worst). Moreover, this means that the expected accuracy score of a random classifier is 192.

Results

For the GloVe Model: We received a mean accuracy score of 137.33 and a median accuracy score of 140.38. In addition, the best fold accuracy score was 104.58, which was received on fold 6. Moreover, the worst fold accuracy score was 174.25, which was received on fold 13 (for the full results distribution, see Fig3). In addition, according to the labeling criteria (previously discussed), 274 sentences (71.4%) were decoded with more success.
For the BERT Model: We received a mean accuracy score of 118.08 and a median accuracy score of 119.33. In addition, the best fold accuracy score was 72.75, which was received on fold 14. Moreover, the worst fold accuracy score was 155.0, which was received on fold 5 (for the full results distribution, see Fig4). In addition, according to the labeling criteria (previously discussed), 301 sentences (78.4%) were decoded with more success.

Discussing the Results

As we can see, both models have excellent performance as attested by the measures mentioned above. However, from the comparison, it appears that the BERT model performed significantly better in every measure. Particularly, notice that both mean and median accuracy scores of the BERT model are approximately 15% better (lower) than those of the GloVe model. Moreover, we noticed that the top 5 most successfully decoded sentences are very different between the two models. The same insight goes for the top 5 least successfully decoded sentences as well. These results further highlight the influence of the word embedding model on the decoder's behavior as seen at the beginning of our analysis.

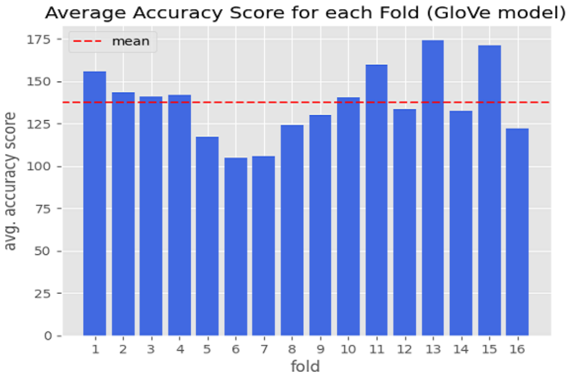

Fig 3

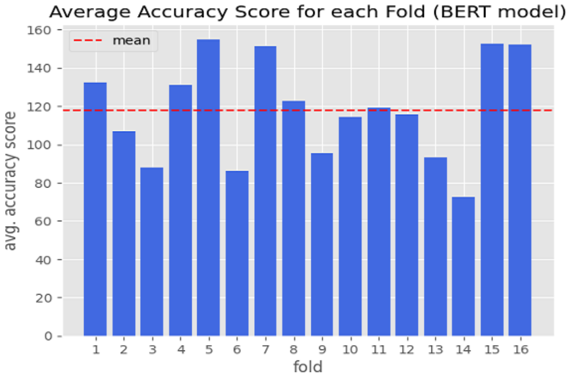

Fig 4

Section 2

In this section, as opposed to the previous ones, we were asked to predict fMRI data from the embedded vectors of sentences. That is, to create a brain-encoder from sentences to fMRI data. In order to achieve our goal, we fit a linear regression model on each of the voxels (a total of 185,866). We used the R^2 and P_{value} metrics to assess each of the regression models' performance. We used the data from experiment 2 (for the same reasons mentioned in the previous section) and performed the task using two different embedding models:

- GloVe model (the embedding model used in the paper).
- BERT model (the contextualized word embedding model we extracted in the previous section).

Results

For the GloVe Model: After finding the linear regression models of all the voxels, we received the following results: The most common R^2 value is approximately 0.8. Moreover, the distribution is symmetrical around this value. Thus, the distribution seems like the Normal distribution with μ=0.8 (for the distribution, see Fig5). Moreover, the vast majority of the regression models are statistically insignificant (84.0%).

For the BERT Model: We noticed that the BERT model embeds words into a 768-dimensional vector space, while we only have 384 sentences in experiment 2 (the experiment with the higher number of sentences between the two). Thus, we have more explaining variables than samples in our linear regression models. As a result, we cannot find the p-value of the model (as required in this section). To overcome this issue, we performed dimensionality reduction (using PCA) on the 768-dimensional embedded word vectors. We chose to reduce their dimension to 300 (same dimension as the GloVe model vectors), which allows us to better compare the results of the two models. After performing PCA (as mentioned above), and finding the linear regression models of all the voxels, we received the following results: The most common R^2 value is approximately 0.8. Moreover, the distribution is symmetrical around this value. Thus, the distribution seems like the Normal distribution with μ=0.8 (for the distribution, see Fig6). Moreover, the vast majority of the regression models are statistically insignificant (71.4%).

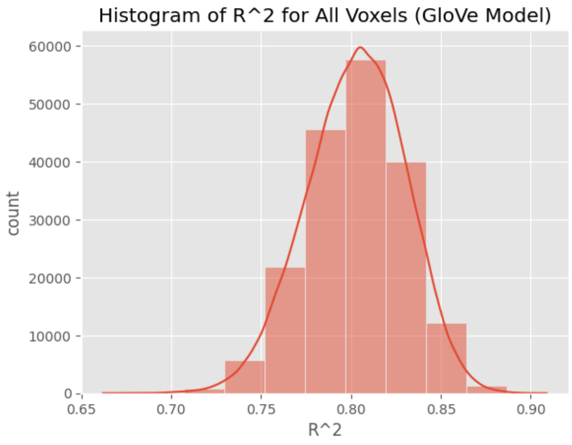

Fig 5

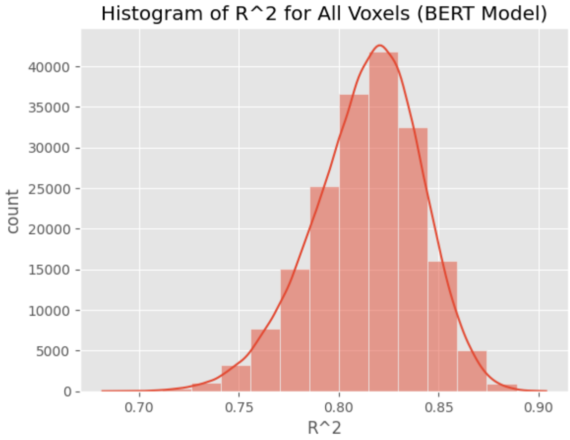

Fig 6

Discussing the Results

First, although we used two entirely different embedding models (static compared to contextualized), the distributions of R^2 values are roughly the same between them (Normal distribution with μ=0.8). Second, in both models, the vast majority of the voxel regression models are insignificant. However, using the BERT embedding model yields approximately twice the number of significant regression models compared to the GloVe model (28.6% significant compared to 16%).

Open-ended Task

In the previous tasks, the analysis was done using only fMRI data from an individual subject. As a natural extension to our prior research, we would now like to perform an analysis using data from multiple subjects. After discussing this idea with a course staff member, we were advised to use the dataset from the paper “Pereira et al., 2018” (the paper Project 2 is based on). This dataset contains brain data sampled from 16 subjects who took part in various experiments (experiments 1-3 as described in the paper). Our open-ended task will be divided into two main analyses: one will focus on the brain decoder, and the other on the brain encoder (as seen in the previous tasks). As mentioned at the beginning, not all 16 subjects took part in all three experiments. However, all subjects participated in experiment 1. Therefore, in this task we focused on data exclusively from that experiment, to gain the highest possible number of subjects we could perform the analysis on. In addition, throughout the analysis we used the GloVe word embedding model (as done in the paper). The analysis done in this part requires an equal number of voxels for each of the subjects (for example, it involves training a model on one subject and evaluating its performance on different subjects). However, this is not the case in our data. Therefore, we cannot use voxels fMRI data for our analysis. On the other hand, we noticed that the number of ROIs (Region of Interest) in all subjects' data is the same (116). Moreover, each subject's data specifies the voxels that compose each ROI. Thus, we chose to use ROIs fMRI data (instead of voxels fMRI data), in order to satisfy the mentioned requirement of the analysis. The ROIs fMRI data (for each of the subjects) was extracted as follows: The raw data of a subject contains voxels fMRI data for each of the three paradigms described in the paper. To attain experiment 1's overall voxels fMRI data which considers all three paradigms, we averaged the voxels data across all three of them (as done in the paper). Then, for each ROI, we averaged the fMRI data of the voxels associated with it, to obtain its fMRI data that we will use in this analysis.

Brain Decoder Analysis

Subjects Cross-Validation Analysis

First, let us define the following terms:

- We will define a "single-subject model" as a model which was trained only on one subject.

- We will define that the single-subject model "corresponding" to subject i, is a model that was trained only on subject i.

Now, we would like to see if a single-subject model performs more successfully on its corresponding subject than on all the others. We hypothesize this is indeed the case (as it is quite reasonable to assume). In order to test our hypothesis, we performed leave-one-out CV on the subjects – for each subject, we trained a brain decoder model on the current subject's data (ROIs fMRI), and evaluated its performance on all the subjects (including the current one). Note, that during this CV process, each single-subject model is also tested on its corresponding subject. In order to avoid evaluating the model on data that it was trained on, for each single-subject model, we performed CV on the data as well. In this CV process, we used 18 folds – each containing 10 concepts. That is, for a given subject and a given test fold, we trained our model on all the other folds (train folds) using only the given subject's data (its corresponding subject), and individually evaluated the model's performance on the test fold of each of the subjects.

Results

In the results analysis we will use the following terms:

- We will define that a single-subject model performs "best" on its corresponding subject if the model receives a better (lower) accuracy score on that subject than on all other subjects.

- We will define that a single-subject model performs "better" on its corresponding subject if the model receives a better (lower) accuracy score on that subject than the mean accuracy score across all other subjects.

It appears that only 1 out of the 16 single-subject models performed best on their corresponding subject. However, 9 out of the 16 single-subject models performed better on their corresponding subject (for the full distribution of each single-subject model, see Fig7).

Discussing the Results

As we can see, our hypothesis is neither completely correct nor wrong. That is, it is very unlikely that a model which was trained on a subject, will perform best on it. However, it is quite likely (slightly more than 50%) that a model will perform better on the subject that it was trained on, than on some other subjects.

Average-Subject Model Analysis

First, let us define the following term:

- We will define an "average-subject model" as a model which was trained on the average data across all subjects.

Now, we would like to see if an average-subject model performs more successfully on a subject, than its corresponding single-subject model. Note that the average data contains some sort of information regarding each of the subjects (as the average across all subjects' data). Thus, for each subject, when evaluating the average-subject model's performance on it, we performed CV on the data (the same data CV mentioned in the previous analysis). We chose to do so in order to avoid evaluating the model on data that it was somewhat trained on.

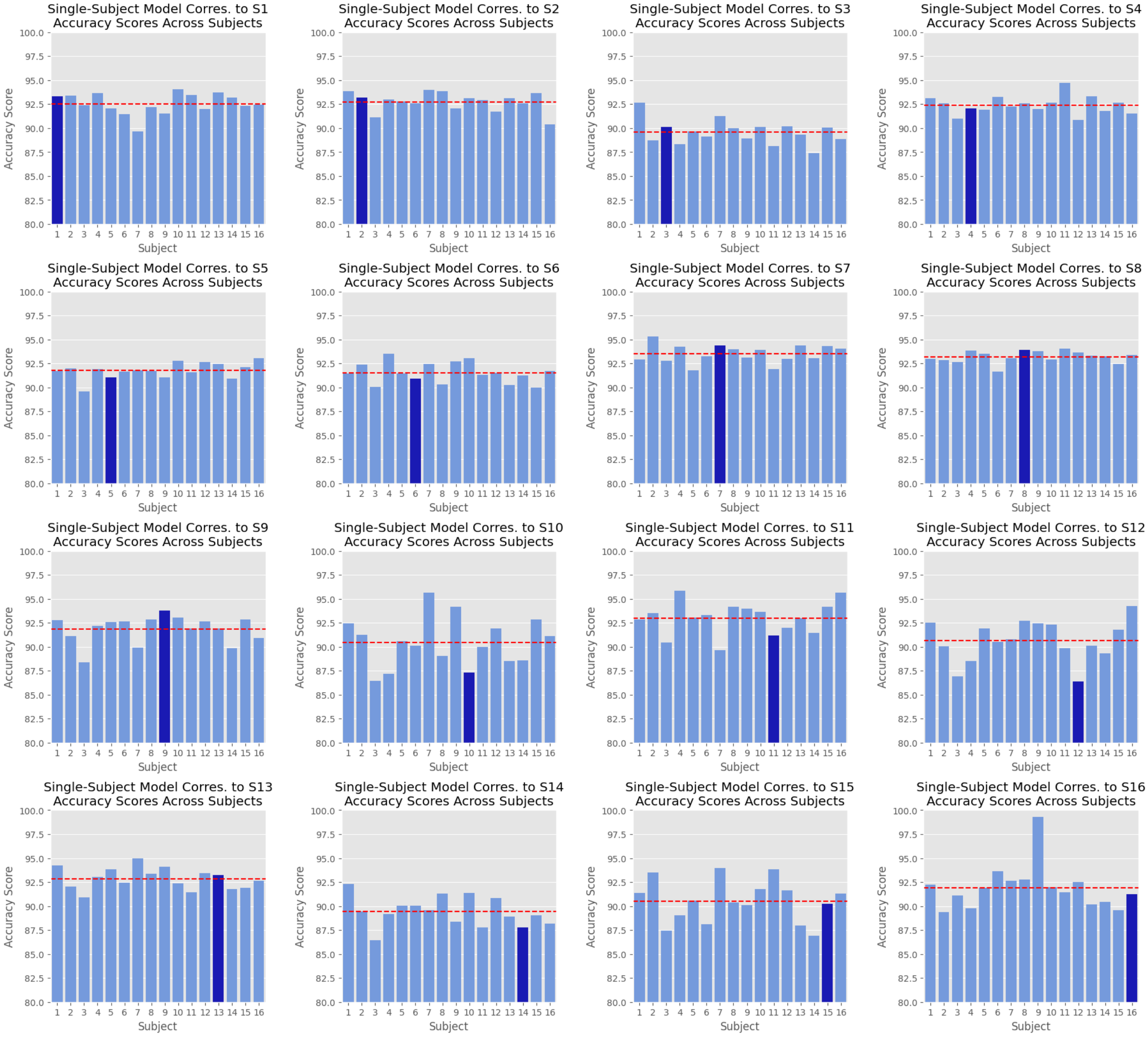

Fig 7

* The bold bar in each plot represents the corresponding subject of the single-subject model.

** The dotted red line represents the mean accuracy score of the single-subject model across all subjects.

Results

It appears that for 15 out of the 16 subjects, the average-subject model received a better (lower) accuracy score on them, than their corresponding single-subject models did (for the full comparison, see Fig8). The average accuracy score improvement across those 15 subjects is 5.22%. Although this improvement is not exceptional, it is also not negligible.

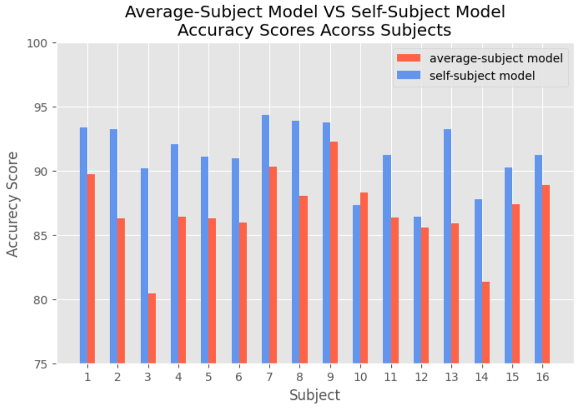

Fig 8

Discussing the Results

As previously mentioned, the average-subject model was somewhat trained on each of the subjects (by training on the average subjects' data). Thus, when tested on a particular subject, it is reasonable to assume that the model will utilize the acquired training knowledge, which is associated with that subject. Moreover, the knowledge acquired on all other subjects may be seen as some sort of regularization, which enhances the model's generalization abilities. Thus, in the next analysis, we would like to see if the results still hold even when the average-subject model does not contain any sort of information regarding the subject that it is tested on.

Average-Subject-Without Model Analysis

First, let us define the following terms:

- We will define a “average-subject-without model” as a model which was trained on the average data across all subjects except a particular subject.

- We will define that the average-subject-without model “corresponding” to subject i is a model that was trained on the average data of all subjects, except subject i.

Thus, we would like to see if for a subject, its corresponding average-subject-without model performs more successfully on it, than its corresponding single-subject model does. As mentioned above, and differently from the previous part, the training data of the average-subject-without model does not contain any sort of information regarding its corresponding subject. Thus, it is not required to perform CV in order to evaluate the model's performance on that particular subject. However, we still chose to perform CV to be able to better compare, for each subject, the performance of its corresponding average-subject-without model and single-subject model, as well as the average-subject model.

Results

It appears that also in this case, for 15 out of the 16 subjects, their corresponding average-subject-without models performed better on them, than their corresponding single-subject models (for the full comparison, see Fig9). The average accuracy score improvement across those 15 subjects is 5.06%. Although this improvement is not exceptional, it is also not negligible. Regarding the average-subject model, first note that the improvement result of both models is almost identical (5.06% compared to 5.22%). Moreover, for each subject, the corresponding average-subject-without model's accuracy score is higher (worse) than the average-subject model by 1.1 at most. Hence, it seems that both models' performance is relatively similar.

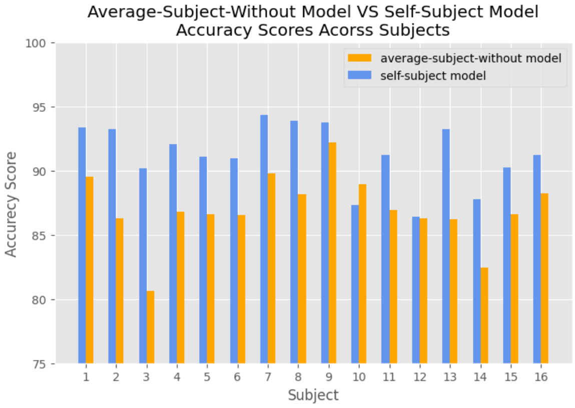

Fig 9

Discussing the Results

As mentioned in the previous part, the average-subject model contains some sort of information regarding each subject. However, in this case, for each subject, its corresponding average-subject-without model was not trained on that subject's data at all (by the definition of the model). This substantial difference makes the results of this analysis much more impressive, especially since both models' performance is approximately the same. Moreover, these results indicate that a model which was trained on a group of random people, can better predict the concept that a person was exposed to, than a model which was trained on his own brain data. Furthermore, these results might suggest the possibility of creating a universal brain decoder model. That is, we can take a large group of random people, train a model on the average brain data of that group, and gain such model. This model will allow us to take a random person and predict the concept that they were exposed to, without the need to fit a model for that person specifically, and even gain greater predictive power.

Brain Encoder Analysis

Inter-Subject Correlation Analysis

In this analysis, we will check for correlation between the degree to which an ROI is mutually active among different subjects (when stimulated with a concept), and the importance of that ROI during concept processing. In order to measure to what degree an ROI is mutually active among the different subjects, we performed Inter-Subject Correlation (ISC): for each ROI, and for a given subject, we calculated the correlation (using Pearson correlation) between that subject's ROI data and the average ROI data of all other subjects. As a result, for each ROI we received 16 correlation scores (one for each subject). Finally, to get the overall correlation score for each ROI, we averaged the 16 scores mentioned above. In order to measure the importance of an ROI during concept processing, we fit a linear regression model for each of the ROIs, and calculated their R^2 value (as done in Section 2 of the Semi-structured Task, except for ROIs fMRI data instead of voxels). Recall that the GloVe model embeds words into a 300-dimensional vector space, while we only have 180 concepts in experiment 1. Thus, we have more explaining variables than samples in our linear regression models. To overcome that, we performed dimensionality reduction (using PCA) and reduced the embedded vectors to 150 dimensions. Note, that the data from experiment 2 contains 384 examples (sentences), which is strictly greater than 300. Thus, we could overcome the mentioned problem and avoid performing PCA (which results in information loss), by using the data from experiment 2. However, only 6 subjects participated in it. In brain data analysis, it is well-known that the number of participants is crucial, due to low signal-to-noise ratio. Therefore, using the data from experiment 2 would have a drastically negative affect on the ISC results. Thus, we chose to keep using the data from experiment 1 (and perform PCA).

Results

After obtaining the results of the tested measures, we performed linear regression to receive the correlation between them, and the statistical significance. We received a linear correlation of 0.54, with very high statistical significance ($PV < 1x10^{-9}$).

Discussing the Results

As we can see, it is statistically significant that there is quite a high positive correlation between the tested measures. In other words, the more an ROI is mutually active among different subjects, the higher the importance of that ROI during concept processing. For visualization purposes, we created a scatter plot (with a linear regression line) where each ROI is represented by a point of form (ISC, R^2) (see Fig10).

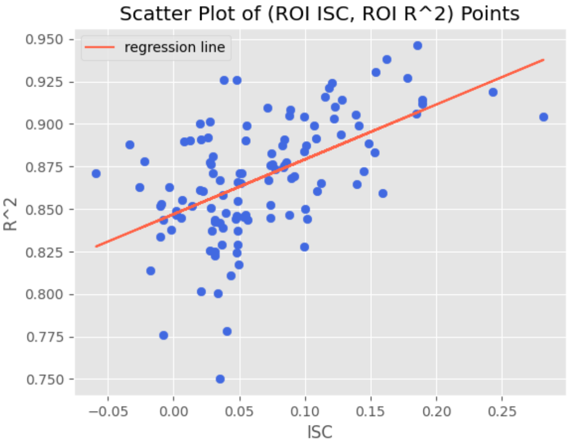

Fig 10

Acknowledgements

This project was based on the article Pereira et al. 2018.

The full Google Colab notebook used for the coding aspect of the project, as well as the data, are provided in the following link.