A Semantic Shared Response Model

Kiran Vodrahalli, Po-Hsuan Chen, Janice Chen, Esther Yong, Christopher Honey, Kenneth A. Norman, Peter J. Ramadge, Sanjeev Arora

Princeton University, Computer Science, Princeton Neuroscience Institute, Electrical Engineering; University of Toronto, Psychology

Objectives

- Given a textual description of a story, what is an accurate way to represent the story context as it changes over time?
- To what extent can we decode semantic descriptions of a story from fMRI readings?

Overview

Several researchers have attempted to find relationships between word featurizations and fMRI activation in the brain. One popular method due to [6] gathers fMRI data across several subjects corresponding to story text. Here we address the multi-view nature of finding meaning in the brain. Our specific goal is to determine if an fMRI shared space can be learned across subjects that correlates well with semantic word embeddings.

We study the Sherlock fMRI dataset [3], which consists of fMRI recordings of 17 people watching the British television program "Sherlock" for 45 minutes. In addition, we use externally annotated, second-level-resolution, English text scene descriptions of the movie.

Using these annotations and the English Wikipedia corpus, we employ unsupervised methods to construct semantic context vectors using global co-occurrence matrix factorization and sparse coding [1, 2]. We then use the unsupervised Shared Response Model (SRM) [4] to construct a shared embedding space across the 17 subjects for eight distinct brain regions of interest (ROI).

Finally, we construct maps between semantic embedding space and the fMRI shared embedding space of our dataset using ridge regression and Procrustes' orthogonal regularization. The models are validated by assessing context vector quality, calculating fMRI reconstruction, and performing binary and scene classification.

Contributions

The setup of this work is novel in a few ways: First, we assume that many subjects viewing the same stimulus will have a consistent internal representations of the events of the movie and model accordingly. This constraint allows us to make use of additional information due to other subjects to both de-noise and find relevant dimensions of brain activity. Second, we are attempting to decode *descriptions* of an audio-visual stimulus while other works which decode text typically use a single-concept stimulus (like a picture of a cat). Therefore in this work, we are truly operating with the **meaning** of both the word descriptions and the fMRI activation.

Constructing Semantic Vectors

To featurize the descriptions of the Sherlock movie, we use the Wikipedia corpus to calculate word co-occurrence values. Weighted singular value decomposition then yields low-rank semantic vectors whose geometry clusters similar words and creates linear algebraic analogy relationships [1]. Recent work has applied sparse coding to these word vectors to get fine-grained 100-dimensional representations of specific word senses called atoms [2]. To construct a single semantic context for each time point, we decompose the associated sentences into (atom, weight) pairs, and run k-means ($k = 4$) on the 100-dimensional atom vectors with weights $> \lambda$. The final context vector is the weighted average of the means.

After generating 100-dimensional context vectors for each time point in the movie, we check the quality of the vectors by finding nearby vectors of fine-grained meaning, which result from the sparse coding step [2]. For instance, consider an example annotation of a scene in Sherlock: *"Donovan looks up at the reporters and continues: 'Preliminary investigations...' Lestrade looks distressed. Donovan continues: '... suggest that this was suicide. We can confirm that this..."*. Nearby word vectors correspond to words like *investigation* (corr. $= 0.78$), *suicide* (corr. $= 0.74$), *CNN* and *Reuters* (corr. $= 0.71$), and *police* (corr. $= 0.70$). The other context vectors have similar quality to this example.

Model Description

There are three components to our model. To construct a shared space for the fMRI data, we use the Shared Response Model (SRM) [4], a probabilistic latent variable model for multisubject fMRI data under a time synchronized stimulus. From each subjects's fMRI view of the movie, SRM learns projections to a shared space that captures semantic aspects of the fMRI response. Specifically, SRM learns orthogonal-column maps W_i such that $\|X_i - W_i S\|_F$ is minimized over $\{W_i\}, S$, where $X_i \in \mathbb{R}^{v \times t}$ is the i^{th} subject's fMRI response (v voxels by t repetition times) and $S \in \mathbb{R}^{k \times t}$ is a feature time-series in a k-dimensional shared space.

$$\underset{W^T W = I, S}{\operatorname{argmin}} \sum_{i=1}^{k} \|X_i - W_i S\|_F \quad (1)$$

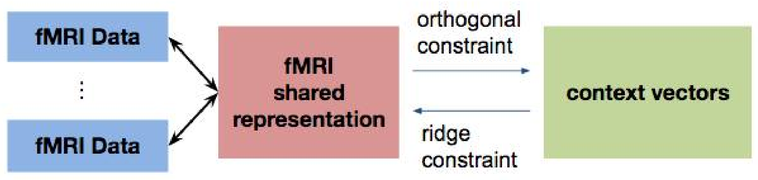

Figure 1: Model Visualization

To learn a map from the semantic context space to the shared fMRI response space, we use ridge regression. For the other direction, we solve the Procrustes problem and learn an orthogonal linear map from shared fMRI response space to the semantic context space, in order to decode.

Experiments

- **Assessing Context Vector Quality:** We examine the time-time correlation matrix of the semantic context vectors, to check that all vectors are completely uncorrelated or completely correlated with each other. There should also be a block-structure along the diagonal signifying different related scenes.

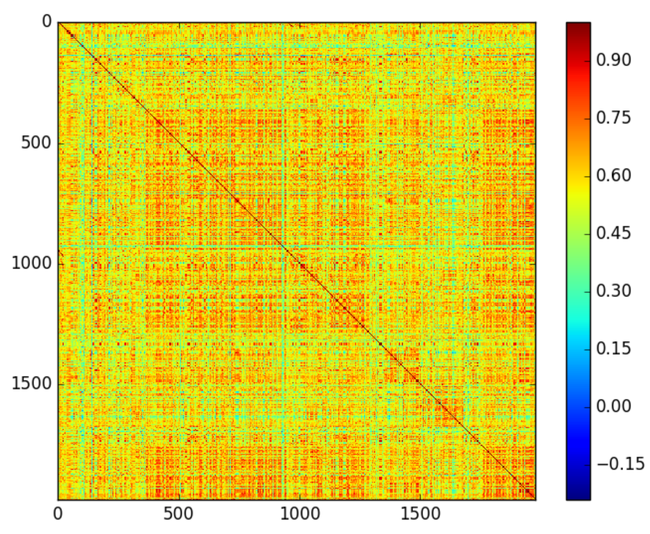

Figure 2: Semantic Vector Time-Time Correlation Matrix Visualization

- **Reconstruction Error:** To align the fMRI and context vectors, we apply ridge regression to learn a linear map from the context vectors to the shared fMRI space. We then predict $\hat{S}$ from the context vectors and compute the Pearson correlation $\langle S, \hat{S} \rangle$. For comparison, we try ridge regression from the context vectors to the individual fMRI responses X_i and compute correlation $\langle X_i, \hat{X}_i \rangle$.
- **Binary Classification:** We segment the time points into 50 evenly sized sections, and train linear maps from fMRI $\rightarrow$ text and text $\rightarrow$ fMRI on 48 of the sections. Then, we use Pearson correlation to match the images of the maps with the held-out values. Success at this task is 50%.
- **Scene Classification:** This task is a harder generalization of binary classification. We use the same segmentation and this time learn a map using only half of the time points. Then, we rank the held-out time points via Pearson correlation and report the average top-1 (4% chance) and top-5 (20% chance) ranks.

We found that 20-dimensional shared fMRI space and 100-dimensional semantic space gave the best overall results, and we report only these.

Results

ROIs [5]	20-dim SRM	raw fMRI
Ventral Language Network	0.15	0.06
Auditory Network	0.11	0.05
DMN (A/B) Network	0.11/0.08	0.04/0.03
Dorsal Language Network	0.10	0.03
Occipital Lobe	0.08	0.04
Early Visual Cortex	0.08	0.04

Table 1: Comparing corr$(\hat{S}, S)$ and avg. corr$(\hat{X}_i, X_i)$

Binary Classification (chance 50%, SRM dim 20, Context dim 100; all errors < 1%

Mask Type	**DMN-A**	**DMN-B**	**Ventral Lang.**	**Dorsal Lang.**
Text -> fMRI (ridge)	**0.83**	**0.76**	**0.8**	**0.79**
Text -> fMRI (procrustes)	0.71	0.68	0.63	0.69
fMRI -> Text (ridge)	0.59	0.6	0.56	0.56
fMRI -> Text (procrustes)	**0.7**	**0.67**	**0.68**	0.61
Mask Type	**Auditory a1+**	**Erez a1**	**Occipital Lobe**	**v1+**
Text -> fMRI (ridge)	0.69	0.71	0.67	0.6
Text -> fMRI (procrustes)	0.6	0.6	0.66	0.61
fMRI -> Text (ridge)	0.57	0.57	0.6	0.57
fMRI -> Text (procrustes)	0.59	0.6	0.66	0.6

Figure 3: Binary Classification Experiment Results

Scene Matching (top-1/ top-5), chance 4%/ 20%, SRM dim 20, Context dim 100; all errors < 1%

Mask Type	**DMN-A**	**DMN-B**	**Ventral Lang.**	**Dorsal Lang.**
Text -> fMRI (Ridge)	**0.26/0.50**	0.17/0.45	**0.18/0.50**	**0.24/0.48**
Text -> fMRI (Procrustes)	0.08/0.38	0.08/0.47	0.11/0.34	0.12/0.43
fMRI-> Text (Ridge)	0.05/0.26	0.12/0.34	0.05/0.21	0.06/0.26
fMRI-> Text (Procrustes)	0.08/0.38	**0.12/0.49**	0.08/0.29	**0.09/0.43**
Mask Type	**Auditory a1+**	**Erez a1**	**Occipital Lobe**	**v1+**
Text -> fMRI (Ridge)	0.07/0.37	0.10/0.4	0.10/0.42	0.08/0.35
Text -> fMRI (Procrustes)	0.05/0.24	0.05/0.25	0.06/0.26	0.04/0.23
fMRI-> Text (Ridge)	0.06/0.24	0.08/0.29	0.09/0.28	0.05/0.24
fMRI-> Text (Procrustes)	0.05/0.24	0.05/0.25	0.07/0.26	0.04/0.23

Figure 4: Scene Classification Experiment Results

Conclusion

The first experiment reveals a significantly lower testing reconstruction error when the fMRI shared space is used, implying that the distributed context embeddings capture some extrinsic notion of meaning that extends beyond a corpus into real-world stimuli.

The binary classification and scene matching experiments demonstrate that using ridge regression for text $\rightarrow$ fMRI and Procrustes for fMRI $\rightarrow$ text yield the best results. Notably, the DMN and language regions outperform the other areas by a fair margin, supporting previous work which suggests that these brain regions encode semantic meaning [3]. DMN-A achieves 83% accuracy at binary classification from text $\rightarrow$ fMRI, outperforming previous work by Mitchell et. al. [6] which achieved 74% accuracy at the same task. DMN-A also achieves 70% accuracy at binary classification from fMRI $\rightarrow$ text.

The top-5 rank scene matching results is 50% for DMN-A (text $\rightarrow$ fMRI) and 49% (fMRI $\rightarrow$ text) for DMN-B. Therefore, we can semi-reliably decode fMRI into semantic space, which is a promising start to decoding thoughts induced by natural stimuli.

References

[1] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. RAND-WALK: A latent variable model approach to word embeddings. *arXiv:1502.03520*, 2015.

[2] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Linear Algebraic Structure of Word Senses, with Applications to Polysemy. 2016.

[3] J. Chen, Y. C. Leong, K. A. Norman, and U. Hasson. Shared experience, shared memory: a common structure for brain activity during naturalistic recall. *bioRxiv preprint*, 2016.

[4] P.-H. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. V. Haxby, and P. J. Ramadge. A Reduced-Dimension fMRI Shared Response Model. *The 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[5] E. Simony, C. J. Honey, J. Chen, O. Lositsky, Y. Yeshurun, and U. Hasson. History dependent dynamical reconfiguration of the default mode network during narrative comprehension. *(in review)*, 2016.

[6] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses.