Mapping Between Natural Movie fMRI Responses and Word-Sequence Representations

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

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

Objectives

- Given a textual description of a movie, what is an accurate way to represent the narrative context as it changes over time?
- To what extent can we map between semantic word representations of the movie and fMRI readings of people watching the movie?

Overview

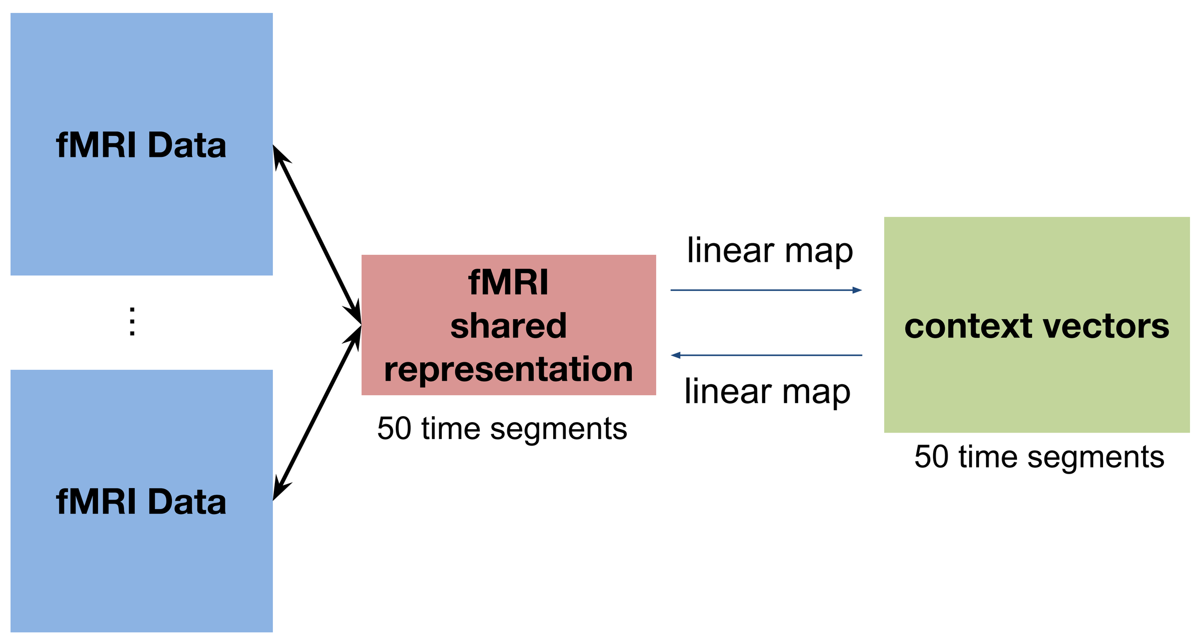
Several researchers have attempted to find relationships between word featurizations and fMRI activation in the brain. One popular method due to [4] gathers fMRI data across several subjects corresponding to story text. We study the Sherlock fMRI dataset [2], which consists of fMRI recordings of 17 people 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. In this poster, we

1. Construct 100-dimensional semantic context vectors for the annotations [1]
2. Apply SRM [3] to construct shared 20-dimensional embedding of originally high-dimensional fMRI subject data
3. Learn linear maps from fMRI $\to$ text and text $\to$ fMRI with ridge regression and the Procrustes problem
4. Evaluate with scene classification (84% over a 20% chance rate) and scene ranking (90% over a 50% chance rate) tasks for five different brain ROIs

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) [3], a probabilistic latent variable model for multisubject fMRI data under a time synchronized stimulus. 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.

To featurize the descriptions of the Sherlock movie, we use the Wikipedia corpus to calculate word co-occurrence values. A matrix factorization objective then yields low-rank semantic vectors whose geometry clusters similar words. In order to combine these representations into vectors for each annotation, each of which is several sentences, we apply a weighted averaging scheme [1]. We learn linear maps from fMRI $\to$ text and text $\to$ fMRI.



Experiments

- **Scene Classification:** We evenly segment the time points into 50 segments and learn a map using the first 25 segments. Then for each predicted held-out segment, we rank via Pearson correlation with the true held-out segments and report the proportion of the time the correct true held-out segment is ranked within the top 5 most correlated segments (20% chance).
- **Scene Ranking:** This task is nearly identical, except we report $1 -$ average normalized rank (1 is highest, 0 is lowest, 0.5 is average random chance).

We compare several pipeline choices in these metrics:

- SRM versus averaging
- Applying a weighted averaging for annotation vectors versus an unweighted average
- Subtracting out the mean of the annotation vectors
- Solving the Procrustes problem (orthogonal constraint) to learn map versus using ridge regression (ℓ_2 constraint).

Results

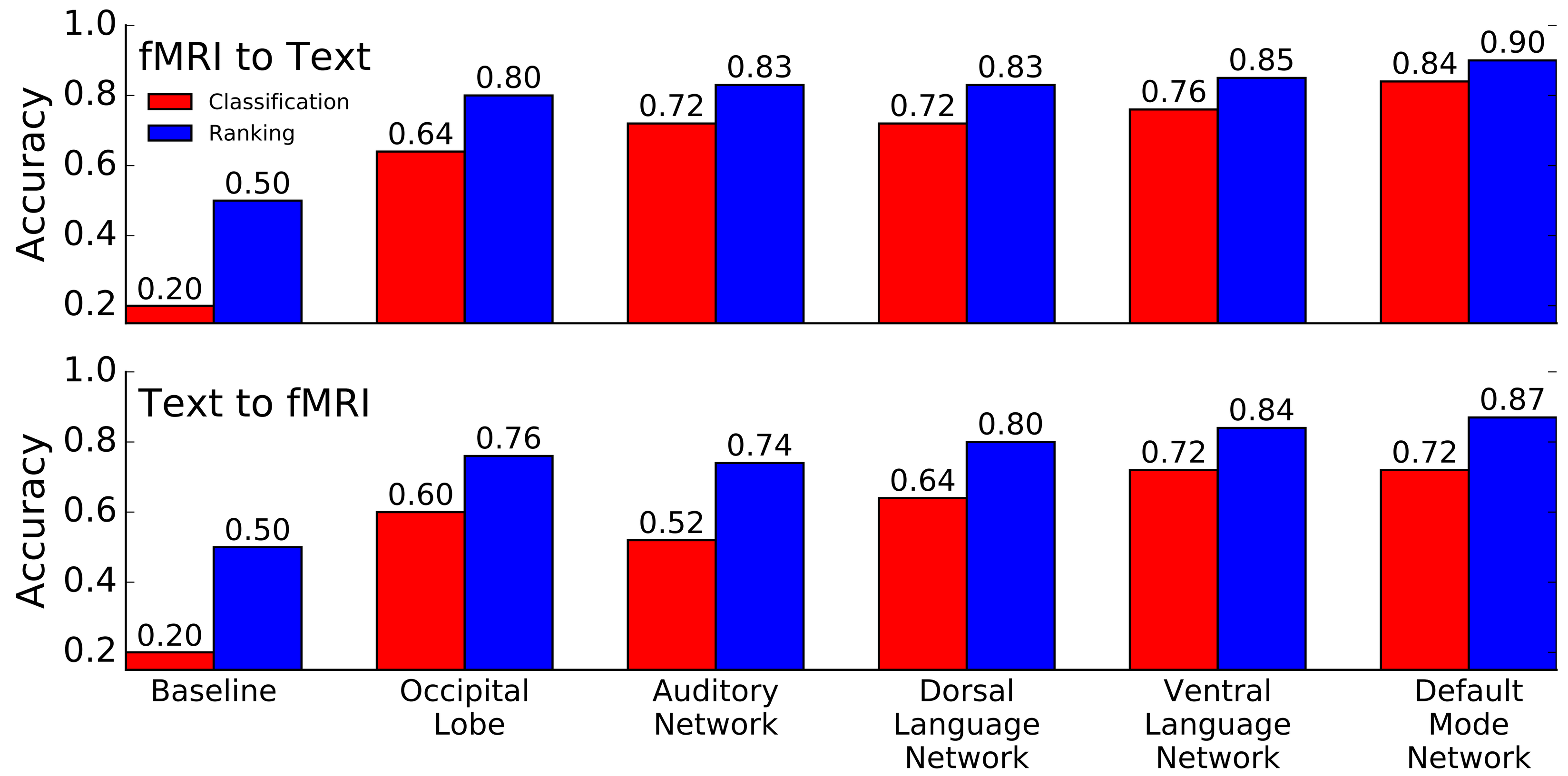


Figure 2: Best Bidirectional Accuracy Scores for Each Brain Region of Interest for both Scene Classification and Ranking (std. err. over different average subsets < 0.01)

Comparison on the Classification Task	fMRI $\to$ Text	Text $\to$ fMRI
20-dim SRM / Avg	1.57 ± 0.10	1.00 ± 0.03
Weighted / Unweighted Semantic Vectors	1.17 ± 0.04	1.06 ± 0.03
Temporal Zero Mean / No Zero Mean	1.09 ± 0.04	1.57 ± 0.11
Procrustes / Ridge	1.42 ± 0.09	0.85 ± 0.06

Table 1: Average Improvement Ratio for Various Comparisons

Discussion

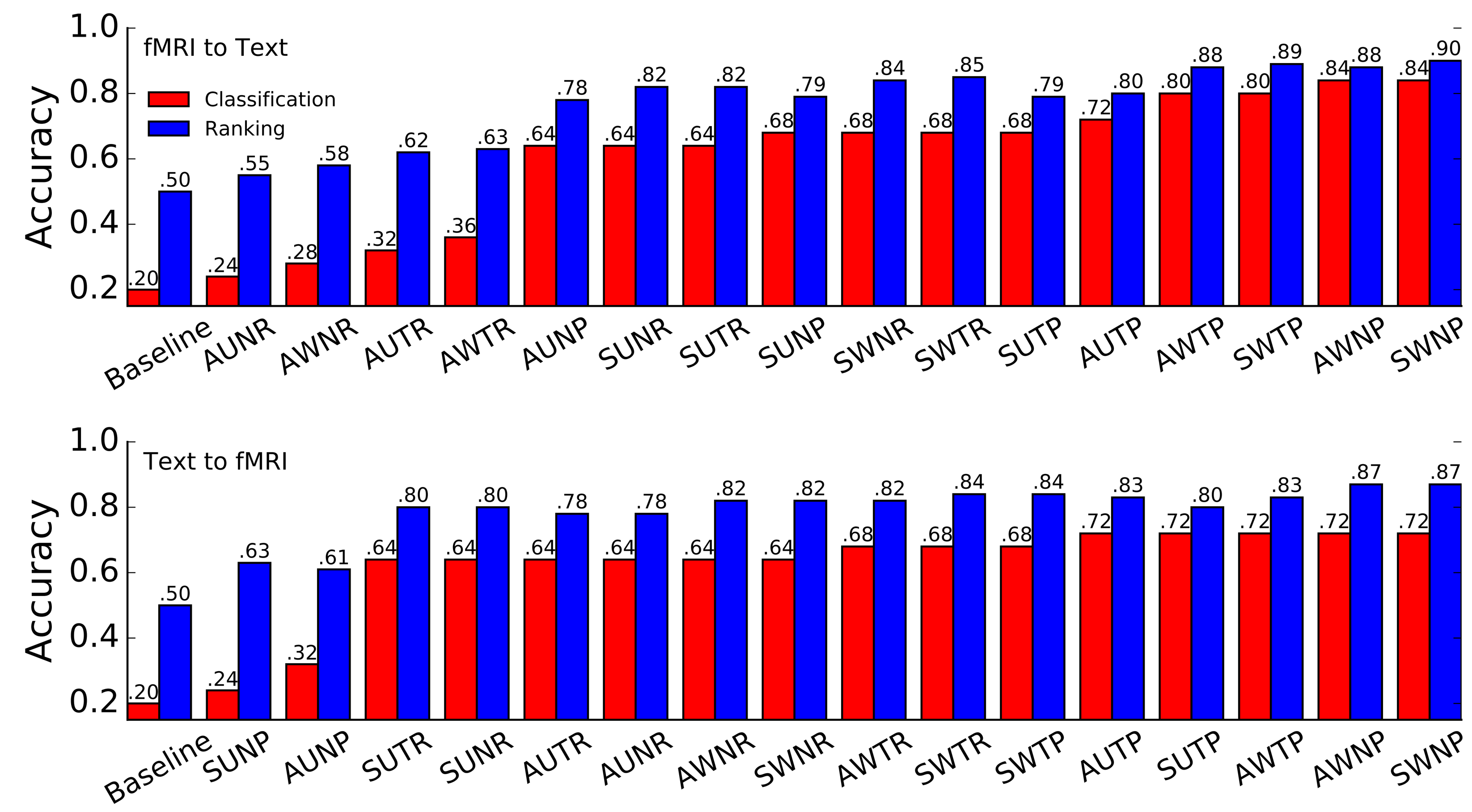


Figure 3: DMN Bidirectional Accuracy Scores for Scene Classification and Ranking. The acronyms stand for combinations of methods, with the following key: S/A = SRM/Average, W/U = Weighting/No Weighted, T/N = Temporal Zero Mean/No Temporal Zero Mean, P/R = Procrustes/Ridge (std. err. over different average subsets < 0.01)

We now list our main findings:

- Top accuracy of 84% on the fMRI $\to$ text scene classification task using the Default Mode Network region of the brain (SRM, weighted, Procrustes, no mean subtraction)
- DMN has best performance for both fMRI $\to$ text and text $\to$ fMRI
- SRM versus averaging improves performance by $1.57\times$ on average, but only considerably improves accuracy over averaging if with averaging the result is bad (by as much as a factor of 2.67)
- Temporal zero mean is the only algorithmic step which seems to make a big difference on average for the text $\to$ fMRI problem, but does not affect the fMRI $\to$ text problem
- Procrustes regularization universally outperforms Ridge regression, on average by a factor of $1.42\times$. Top six methods use Procrustes.
- Weighted combination of word vectors proves average of $1.17\times$ improvement for fMRI $\to$ text; top three methods use weighted word vectors.

References

[1] S. Arora, Y. Liang, and T. Ma.
A Simple but Tough-to-Beat Basline for Sentence Embeddings.
preprint, 2016.

[2] 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.

[3] 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.

[4] L. Wehbe, A. Vaswani, K. Knight, and T. Mitchell.
Aligning context-based statistical models of language with brain activity during reading.
pages 233–243, 2014.