A temporal decay model for mapping between fMRI and natural language annotations

Kiran Vodrahalli, Cathy Chen, Viola Mocz, Christopher Baldassano, Uri Hasson, Sanjeev Arora, Kenneth A. Norman

Princeton Computer Science, Princeton Neuroscience Institute

Objectives

- What is a good way to represent brain signals due to real world stimula as they change over time? Particularly, what models allow for the identification of different time scales in the brain which can be learned?
- Does using temporal information over a window of time improve the performance of linear maps going between fMRI space and a textual meaning space?

Overview

Several researchers have attempted to find relationships between word featurizations and fMRI activation in the brain [4, 5, 7]. 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 work, we identify a **temporal decay** model for combining fMRI information at multiple timepoints which is both interpretable and successful in a prediction task. In this poster, we follow the work of [6] and

1. Construct 100-dimensional semantic context vectors for the annotations [1] as in [6]
2. Apply SRM [3] to construct shared 20-dimensional embedding of originally high-dimensional fMRI subject data as in [6]
3. Learn temporal weight parameters as part of fMRI $\rightarrow$ text and text $\rightarrow$ fMRI regression problems

Model Description ([6])

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 $\rightarrow$ text and text $\rightarrow$ fMRI, using multiple timesteps.

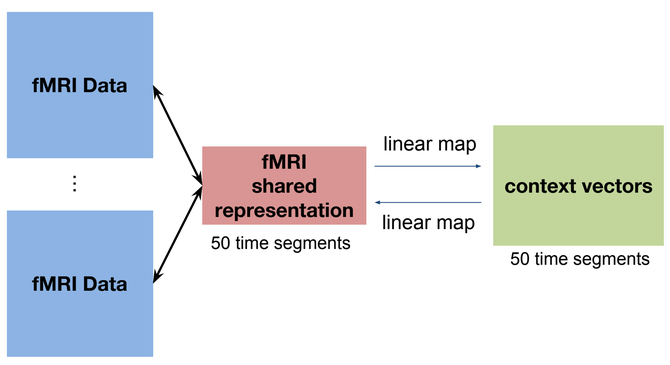

Figure 1: Model Visualization

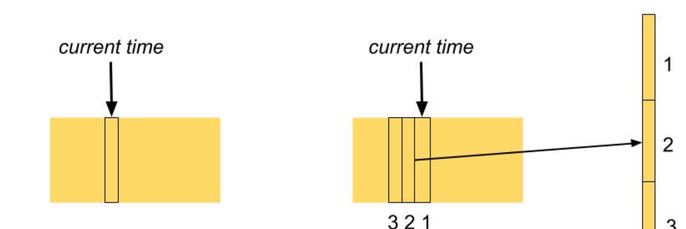

Figure 2: Multiple Timesteps: A visualization of $\hat{X} \in \mathbb{R}^{n*(k+1)}$.

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 1 most correlated segments (4% 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).

Temporal Weighting Models

In fMRI $\rightarrow$ Text tasks, $X \in \mathbb{R}^{n \times T}$ = fMRI space and $Y \in \mathbb{R}^{m \times T}$ = Text space, where n and m are the dimensions of embeddings in these spaces and T is the number of timepoints. In Text $\rightarrow$ fMRI, the variables are flipped.

No Previous Timesteps This model is the simplest, and uses no previous timepoint information. We learn $W \in \mathbb{R}^{m \times n}$:

$$WX = Y \tag{1}$$

Weighted Average In this model, we assume that the past timesteps are important, and that the ideal representation is a linear combination of past timesteps. We learn $W \in \mathbb{R}^{m \times n}$ and convolution matrix $\Phi \in \mathbb{R}^{T \times T}$ such that

$$WX\Phi = Y \tag{2}$$

where column Φ_t is defined by $\Phi_t(i) = \phi_j$ if $i = t - j$, 1 if $i = t$, and 0 otherwise. However in practice, the optimal weights set $\Phi = I$, which reduces to the no previous timesteps case.

Full Temporal Model [6] We learn $\hat{W} \in \mathbb{R}^{m \times n*(k+1)}$ such that

$$\hat{W}\hat{X} = Y \text{ where } \hat{X} \in \mathbb{R}^{n*(k+1) \times T} \tag{3}$$

For a visualization of $\hat{X}$, see Figure 2. The key feature of this model is that weighting parameters for every feature at every timestep in the linear regression model are learned.

Temporal Decay Model We now specify n different decay weights $\lambda = [\lambda_1, \cdots, \lambda_n]$ for each of the fMRI features in the fMRI $\rightarrow$ Text setting (likewise m in Text $\rightarrow$ fMRI). We formulate the problem setting

$$WC_k\hat{X} = Y \tag{4}$$

where $W \in \mathbb{R}^{m \times n}, C_k \in \mathbb{R}^{n \times n*(k+1)}, \hat{X} \in \mathbb{R}^{n*(k+1) \times T}$, and $Y \in \mathbb{R}^{m \times T}$. We define a concatenation C_k of k diagonal matrices Γ_j

$$C_k = [\Gamma_0, \Gamma_1, \cdots, \Gamma_k] \text{ where}$$
$$\Gamma_j(i,i) = \frac{e^{j\lambda_i}}{Z_i} \text{ and } Z_i = \sum_{j^*=t}^{t-k} e^{(t-j^*)\lambda_i} \text{ (normalization)} \tag{5}$$

In this setting, we learn a unique decay parameter λ_i for every feature, which controls the weight at each time step for each feature via an exponential decay function.

Results

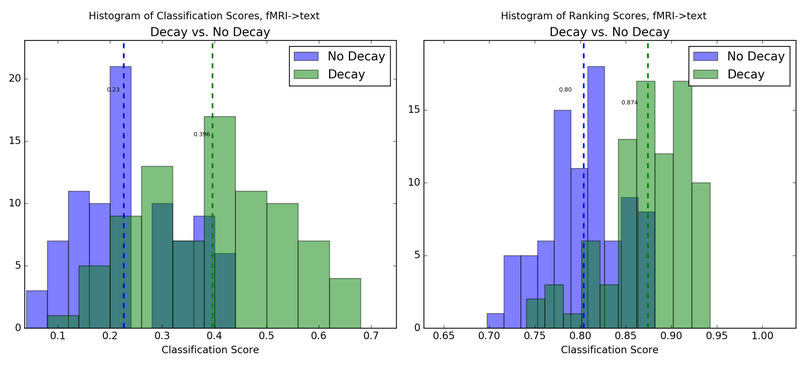

Figure 3: Comparing the performance of using previous timesteps with temporal decay to using no previous timestep information for the fMRI $\rightarrow$ Text classification (chance rate 4%) and ranking (chance rate 50%) tasks. We use the DMN region for this figure. The histogram is over all possible selections of parameters, as explored in [6], including regression model (Procrustes or ridge regression), dimension reduction method (PCA, SRM, SRM-ICA), and number of previous timesteps (ranging from $0-30$) used. Using around $5-8$ previous timesteps is optimal.

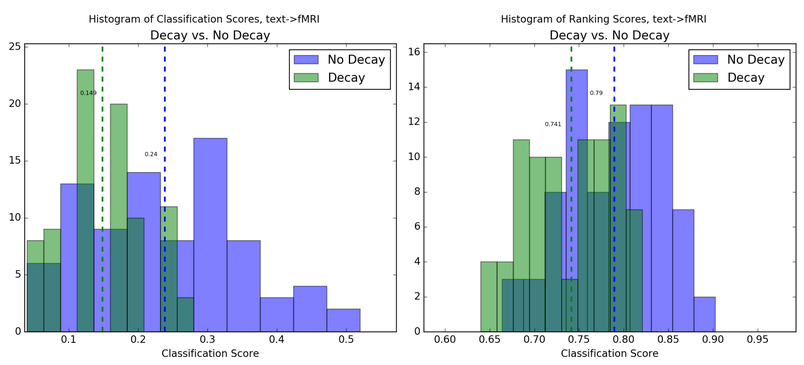

Figure 4: Comparing the performance of using previous timesteps with temporal decay to using no previous timestep information for the Text $\rightarrow$ fMRI classification (chance rate 4%) and ranking (chance rate 50%) tasks. We use the DMN region for this figure. The histogram is over all possible selections of parameters, as explored in [6], including regression model (Procrustes or ridge regression), dimension reduction method (PCA, SRM, SRM-ICA), and number of previous timesteps (ranging from $0-30$) used. Using no previous timesteps is optimal.

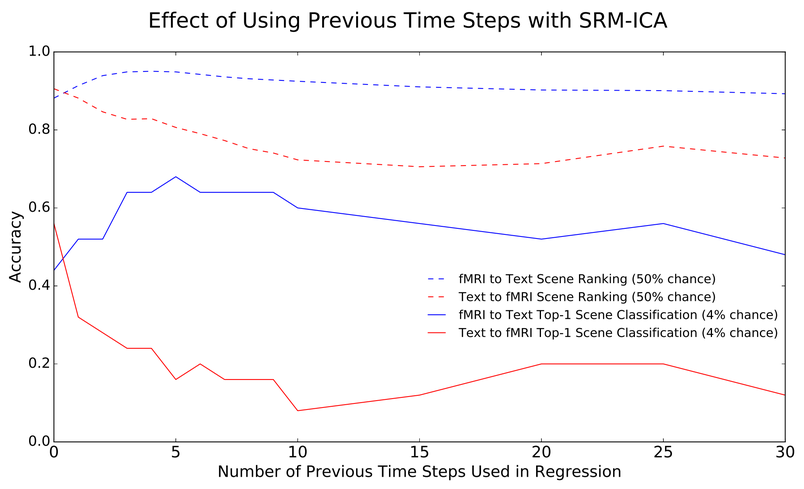

Figure 5: Performance over number of previous timepoints used for the DMN region on all four tasks.

Discussion

	fMRI $\rightarrow$ Text	Text $\rightarrow$ fMRI
Full Temporal ([6])	64%	20%
Temporal Decay (this work)	64%	28%
No Previous Timesteps	44%	56%

Table 1: Comparing the best performances on the classification task (4% chance rate) for our three temporal models in the DMN region. We note that temporal decay does at least as well as the full temporal model in both settings, though it seems to hurt in the Text $\rightarrow$ fMRI case.

We now list our main conclusions:

- The temporal decay model performs at least as well as the full temporal model, but with many fewer parameters to learn (only $(m+1)n$ compared to $mn(k+1)$). In the temporal decay model, we learn a single decay parameter for each feature. It seems critical that we allow for variation in decay weights across features, while it is less important to control exactly how variation in weights occurs across time: Assuming an exponential decay model suffices.
- Building on this point, the weighted decay model (with $mn + k$ parameters) does not work at all. In the weighted average model, we learned a single weight parameter for each timepoint. Variation over weights in time without allowing variation over features is useless.
- In the fMRI $\rightarrow$ Text case, the decay weights correspond to measures of timescale for each of the features, leading us to a more neuroscientifically interpretable model, as well as a way for potential future approaches to characterize time scales in different parts of the brain.
- fMRI $\rightarrow$ Text performs better in raw scores than Text $\rightarrow$ fMRI. This result may be due to the fact that the semantic annotation representations tend to be more stationary, implying that Text $\rightarrow$ fMRI is a one-to-many problem, which is considerably harder to solve than the many-to-one problem of fMRI $\rightarrow$ Text.
- fMRI $\rightarrow$ Text benefits from using previous timepoint information, while Text $\rightarrow$ fMRI does not. Stationarity of the text representations may lead to unneccessary additional parameters in the model (no new information is added by considering the previous time points, since they are similar enough). This result may reflect a deficiency in the annotation data: Humans who summarize the annotation already perform some aggregation over timescales as a result of their natural ability to understand narratives. Future work should edit the annotations to be more fine-grained.

Acknowledgements

The dataset is online [2] and the code used in this paper is available on GitHub. Additionally, we note that we used `http://brainiak.org/` for some of the implementations of algorithms used in this paper. This work was funded by a grant from the Intel Corporation, NIMH R01MH112357 awarded to U. Hasson and K. Norman; NIH grants R01-MH094480 and 2T32MH065214-11; NSF grants CCF-1527371, DMS-1317308, Simons Investigator Award, Simons Collaboration Grant, and ONRN00014- 16-1-2329 awarded to S. Arora.

References

[1] S. Arora, Y. Liang, and T. Ma.
A Simple but Tough-to-Beat Baseline for Sentence Embeddings.
International Conference on Learning Representations (ICLR) 2017, 2017.

[2] J. Chen, Y. C. Leong, C. J. Honey, C. H. Yong, K. A. Norman, and U. Hasson.
Shared memories reveal shared structure in neural activity across individuals.
Nature Neuroscience, 20:115–125, 2017.

[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] A. G. Huth, W. A. deHeer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant.
Natural speech reveals the semantic maps that tile human cerebral cortex.
Nature, 532:453–458, 2016.

[5] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just.
Predicting Human Brain Activity Associated with the Meanings of Nouns.
Science, 320:1191–1194, 2008.

[6] K. Vodrahalli, P.-H. Chen, Y. Liang, C. Baldassano, E. Yong, C. Honey, U. Hasson, P. Ramadge, K. Norman, and S. Arora.
Mapping between fmri responses to movies and their natural language annotations.
NeuroImage, 2017.

[7] 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.
PLOS ONE, 9, 2014.