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

Kiran Vodrahalli (knv@princeton.edu)
Computer Science, 35 Olden Street
Princeton, NJ 08544 United States

Cathy Chen (cc27@princeton.edu)

Viola Mocz (vmocz@princeton.edu)

Christopher Baldassano (cbaldassano@princeton.edu)

Uri Hasson (hasson@princeton.edu)

Sanjeev Arora (arora@cs.princeton.edu)

Kenneth A. Norman (knorman@princeton.edu)

Keywords: fMRI decoding; semantic annotations; multimodal dataset; natural stimuli, natural language understanding; word sequence embeddings; temporal dynamics

Introduction

Recent work has provided convincing evidence that fMRI readings from human subjects can be related to semantics of presented stimuli. Such experiments consist of finding (1) low-dimensional representations of the fMRI signals, and (2) low-dimensional semantic representations of the external stimulus. These tasks build upon work in machine learning.

The earliest work concerned simple settings with carefully controlled stimuli, such as subjects being presented (visually or auditorily) with one of a set of carefully selected words as in Mitchell et al. (2008). Several recent papers attempt variants of this goal on more natural stimuli like audio stories as in Huth, deHeer, Griffiths, Theunissen, and Gallant (2016). An even more recent prior work by Vodrahalli et al. (2017) explores methodology for improving the performance of text-fMRI and fMRI-text maps at scene classification and ranking tasks on a natural movie stimulus with semantic annotations dataset (J. Chen et al., 2017). In particular, they find that using the Shared Response Model (SRM) from P.-H. Chen et al. (2015) to aggregate fMRI data from multiple subjects is superior to applying standard PCA for producing low-dimensional representations. Additionally, they show a sentence embedding technique adapted from the natural language processing (NLP) literature due to Arora, Liang, and Ma (2017) produces useful semantic vector representations of the annotations. Finally, they demonstrate that using previous timepoint information in the setting of predicting text from fMRI is very helpful, though the resulting maps are not very interpretable.

In this short paper, we present novel tweaks to the temporal dynamics methodology of Vodrahalli et al. (2017) which result in great improvements in interpretability with only small penalties in overall accuracy.

Datasets and Tasks

In this work, we study the Sherlock annotation dataset (J. Chen et al., 2017; Vodrahalli et al., 2017). The Sherlock dataset consists of fMRI recordings of 16 people watching the British television program "Sherlock" for 50 minutes broken into 1973 TRs, where each TR is 1.5 seconds of film. In this work, we focus on the default mode network (DMN) region, a brain area known for its relevance to narrative understanding Vodrahalli et al. (2017). As a proxy for the semantics of the movie, we use externally annotated English text scene annotations of the program (average annotation length 15 words per TR). We account for hemodynamic response by shifting the fMRI signal 3 TRs back, the approximate delay length.

We evaluate our ability to map between fMRI and text annotation representations with the **scene classification** task. Given a map from fMRI to text space or its inverse, we apply the map to heldout timepoints and form "scene chunks" which partition the test TRs. Scene classification measures the accuracy of Pearson correlation between predicted and true scene chunks in determining the identity of a held-out test chunk. Random guessing gives a chance rate of $1/$num. of scenes. In our experiments, the chance rate is 4%.

Methodology

We apply the same primary techniques as Vodrahalli et al. (2017) to achieve state-of-the-art fMRI-annotation mapping accuracy, with the exception of the temporal representation.

Interpretable Temporal Dynamics Model

When predicting fMRI $\rightarrow$ Text and Text $\rightarrow$ fMRI, we would like to figure out how to use previous time steps in our linear maps. Let X denote fMRI data and Y denote text data. For the fMRI $\rightarrow$ Text task, we would like to find a map from some representation of the fMRI data to some representation of the Text data. Vodrahalli et al. (2017) learn a unique weight for every feature $1, \cdots, n$ of X for every single timepoint in the previous k time points. This **full temporal** model is simply expressed as

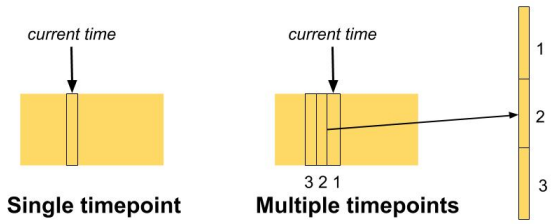

Figure 1: *By adding previous timesteps, we transform the base space into representations of the dynamics of the data.*

learning $\hat{W} \in \mathbb{R}^{m \times n*(k+1)}$ such that we have $\hat{W}\hat{X} = Y$, where $\hat{X} \in \mathbb{R}^{n*(k+1) \times T}$.

A simpler, smaller, more interpretable model might imagine that previous timestep information could be represented as a weighted aggregation over k previous timesteps. However, empirically, weighted aggregation in this form hurts performance. We thus relax the requirement that we have a single weight for each column, and allow different columns to have different weights as in Vodrahalli et al. (2017). However, we want to enforce additional assumptions on the parameters so that the model is more interpretable. We can compromise by defining the **temporal decay model**, which assumes that the weight parameters decay exponentially over the past k timesteps, at a different rate for each representation feature.

The neuroscientific motivation behind the assumption that there may be different rates of decay for different fMRI features comes from the notion that different parts of the brain operate over different time scales: The neurons in some parts of the brain fire a lot more rapidly and react to quickly changing stimuli, while other parts of the brain fire much more occasionally and change according to real world stimuli which occur at longer time scales.

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. We formulate the problem setting $WC_k\hat{X} = Y$, 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 $C_k = [\Gamma_0, \Gamma_1, \cdots, \Gamma_k]$ where $\Gamma_j(i,i) = \frac{e^{j\lambda_i}}{Z_i}$ and $\Gamma_j(i,h) = 0$ when $i \neq h$. Here, $Z_i = \sum_{j^*=t}^{t-k} e^{(t-j^*)\lambda_i}$ normalizes each row.

Results

The scene classification performances for the fMRI $\rightarrow$ Text setting in the DMN are 64% (both full temporal Vodrahalli et al. (2017) and temporal decay models) and 44% (no previous timesteps). In the Text $\rightarrow$ fMRI setting for the DMN, the classification performances are 20% (full temporal model), 28% (temporal decay model), and 56% (no previous timesteps).

These results demonstrate that we can replace the model of Vodrahalli et al. (2017) with a more interpretable model with no loss in the fMRI $\rightarrow$ Text setting, at least for the DMN region. In the Text $\rightarrow$ fMRI setting, we see that the interpretable model improves upon the full temporal model slightly, though both temporal models are worse off compared to the no-previous-timestep model. The same conclusions from Vodrahalli et al. (2017) with respect to shared space dimension reduction, word embeddings, and brain ROI performances hold when we replace the temporal dynamics model.

Conclusion and Future Work

In this work, we presented an interpretable temporal dynamics model which improves maps from fMRI to Text. It remains to explain why Text to fMRI does not perform as well, a problem noted by Vodrahalli et al. (2017) as well. We believe the central reason is due to the relatively high correlation between the semantic representations compared to the correlation between fMRI states. As a result, we perform a one-to-many task when we attempt Text to fMRI, which is more difficult. Future work should thus further decorrelate the text representations.

Acknowledgments

The dataset is online (J. Chen et al., 2017) and the code used in this paper will be made 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

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

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*, 115-125.

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

Huth, A. G., deHeer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*, 453-458.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, *320*, 1191-1194.

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Yong, E., Honey, C., ... Arora, S. (2017). Mapping between fmri responses to movies and their natural language annotations. *arXiv preprint*. Retrieved from `https://arxiv.org/pdf/1610.03914.pdf`