Mapping Between Natural Movie fMRI Responses and Word-Sequence Representations

Kiran Vodrahalli[1], Po-Hsuan Chen[1], Yingyu Liang[1],
Janice Chen[1], Esther Yong[3], Christopher Honey[2],
Peter Ramadge[1], Ken Norman[1], and Sanjeev Arora[1]

[1]Princeton University, [2]Johns Hopkins University, [3]University of Toronto

Abstract. This work provides support for the notion that distributional methods of representing word meaning from computational linguistics are useful for capturing neural correlates of real life multi-sensory stimuli, where the stimuli —in this case, a movie being watched by the human subjects— have been given text annotations. We present an approach to combining sequences of word vectors into a single vector. We also identify a semantically-relevant low-dimensional shared representation of fMRI response in an unsupervised fashion by using views of multiple subjects watching the same natural movie stimulus. Learned orthogonal linear maps between the fMRI and semantic representations allow us to successfully transform fMRI data generated by a natural movie stimulus into semantic vectors representing textual descriptions of the movie. We succeed at a scene classification task with 76% accuracy, over a 20% chance rate. When we selected five brain regions-of-interest (ROIs) and learned distinct maps from these ROIs to the text representations, the Default Mode Network (DMN) supported the highest level of decoding performance.

1 Introduction

Several researchers have tried to relate fMRI readings from multiple subjects to the semantics of the stimulus (text, movies etc.) being presented to the subject, in order to understand the neural correlates of meaning representations. To accomplish this task, we first need a suitable representation of the semantics of the signal. Previous studies have attempted to solve this problem many different ways ([1, 2, 3]). This paper contributes to this literature by applying recent work in computational linguistics for constructing geometric representations of meaning for small text portions (in this case, text annotations of the events unfolding in a movie). This representation of text meaning strongly correlates with fMRI response, allowing bidirectional matching.

We study the **Sherlock** fMRI dataset ([4]), which consists of fMRI recordings of 16 people watching the British television program "Sherlock" for 45 minutes (1973 TRs, where each TR is 1.5 seconds of film). In addition, we use externally annotated, sub-second-resolution, English text scene annotations of the program. For each TR, we have an fMRI recording of each of the subjects' brains as well as a textual description of the events occurring in the movie. Using the Sherlock annotations and the Wikipedia corpus, we employ unsupervised methods to construct semantic context vectors using global co-occurrence matrix factorization ([5, 6]). We employ an averaging scheme ([7]) to combine the individual word vectors into a single semantic vector for each annotation. Next, we use the unsupervised Shared Response Model (SRM) ([8]) to construct a shared embedding space across the 16 subjects for five distinct brain regions of interest (ROI), including the Default Mode Network (DMN). Our interest in the DMN stems from previous work ([9]) that demonstrated that the DMN is related to processing narrative flow. Finally, we construct linear maps from the fMRI shared embedding space of our dataset to the semantic space of the annotations, and vice versa. The models are then validated with two experiments: fMRI scene classification and fMRI scene ranking. Our experiments employ a $50-50$ training-testing split on the first and second halves of the movie.

Our main results are (i) showing that fMRI responses from multiple individuals can be effectively combined using SRM to improve the matching accuracy between the fMRI and the text annotation (Table 1), (ii) presenting a suitable weighted average that improves upon the simple average (Table 1), (iii) improving the distributional representation of text meaning by subtracting the average activation, which significantly improves matching accuracy (Table 1), and (iv) comparing several different brain regions of interest (ROIs), including the auditory region, the visual processing area, ventral and dorsal language regions and the default mode network (DMN) (Figure 1).

The present work is similar in some ways to that of [10], which mapped text embeddings from narrative stimuli to fMRI data. The main differences are our novel semantic vector embeddings and application of SRM, as well as the fact that we successfully learn maps from both fMRI to text and text to fMRI.

2 Methods

2.1 Constructing and Aggregating Semantic Vectors

In order to represent words in a vector space, we take advantage of the distributional properties of words in a large corpus - namely, English Wikipedia. We train vectors as described in [6], which frames the problem of estimating word vectors in the language of generative models. The central assumption is the probability model for a word w given a context c, where the context represents a small window of words in the corpus. We have $\mathcal{P}\left[w|c\right] = \frac{1}{Z_c}\exp(v_w^T c)$ where v_w represents the vector for a given word and Z_c is the value of the partition function which normalizes the distribution. The idea is that the context c represents the subject matter of the text at a given point in time.

Now for every 1.5-second time-point in our Sherlock movie, we have a textual description of what is happening in the movie: actions, dialogue, and so on. This annotation is typically a few sentences long. We can think of each annotation as the current context of the movie narrative. Since we have a word vector for each word in the annotation, we can do the simplest thing possible, which is just averaging all the word vectors to come up with a single annotation vector for each time point. This approach coincides with the maximum likelihood estimate for the context c given some small window of words, given the model from [6]. We will call these representations the **unweighted** annotation vectors.

However, we can try to weight the importance of each word by the amount of specific information it contains. Given our approach to constructing the word vectors, the vectors of words which occur with much greater frequency in the original corpus may inherently contain less information, since these words are in some sense uniform with respect to the whole word distribution. We therefore would like to weight more frequent words less. Following [7], we modify the generative model from [6]: For a word w given context c, the probability of a word w given context c is

$$\mathcal{P}\left[w|c\right] = \alpha\mathcal{P}\left[w\right] + (1-\alpha)\tfrac{\exp(v_w^T c)}{Z_c} \tag{1}$$

where Z_c normalizes the distribution and $\alpha \in [0,1]$. We can think of this model as a convex combination of the probability of a word w appearing not conditioned on the context c and the probability of a word w appearing conditioned on the context c. This setup allows us to take into account the low-information high-frequency words, and remove their importance during our estimation of c.

The MLE estimate of the context vector c in this modified objective is

$$v_{\text{annotation}} = \sum_{\text{word}\in\text{annotation}} \frac{\beta}{\beta + p_{\text{word}}} \cdot v_{\text{word}} \tag{2}$$

where $\beta := \frac{1-\alpha}{\alpha Z}$. Typically, we choose α such that $\beta \approx 10^{-4}$. We call these representations the **weighted** annotation vectors.

Now we have T annotation vectors, one for each time step. On the training portion of the data (the first half of the movie), we calculate an average annotation vector and subtract it from all data. We call this step **temporal zero mean** and assume that the average annotation vector is invariant.

2.2 Shared Response Model for Multi-Subject fMRI

The Shared Response Model (SRM) [8] is 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 N orthogonal-column maps W_i such that $\|X_i - W_i S\|_F$ is minimized over $\{W_i\}_{i=1}^N, S$, 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. In this paper, $k = 20$ since low-rank SVD with 20 dimensions captures 90% of the variance of the original fMRI matrices.

Note that for testing, the learned W_i allow us to project unseen fMRI data into the shared space via $W_i^T X_i^{\text{test}}$ since W_i has orthogonal columns. In this work, we compare the average SRM-shared space projections $\frac{1}{N}\sum_{i=1}^N W_i^T X_i^{\text{test}}$ to simple averaging of the different responses $\frac{1}{N}\sum_{i=1}^N X_i^{\text{test}}$.

We identify five distinct brain regions of interest (ROIs) which we treat completely separately. That is, we first apply ROI masks to the whole-brain data and then learn SRM-representations for each of these ROIs separately.

2.3 Learning Linear Maps

Our approach to predicting semantic vectors from fMRI vectors and vice versa is simply linear regression with two kinds of regularization. Letting $X \in \mathbb{R}^{v \times T}$ represent the fMRI data matrix (either average or SRM) for a specific ROI and $Y \in \mathbb{R}^{100 \times T}$ represent the annotation vectors, our main approach is given by solving the Procrustes problem $\min_{\Omega} \|Y - \Omega X\|_2^2$ with orthogonal columns constraint $\Omega^T \Omega = I_{v \times v}$. Thus, we learn a matrix $\Omega \in \mathbb{R}^{100 \times v}$ as a map from $X \rightarrow Y$, decoding fMRI vectors into semantic space. Our other approach is given by the ridge regression problem $\min_{\omega_j} \|y_j - \omega_j^T X\|_2^2 + \|\omega_j\|_2^2$ where $j \in [1, 100]$ for each word vector dimension. Putting the ω_j together forms $\Omega \in \mathbb{R}^{100 \times v}$ as before, with the orthogonality constraint replaced by a row-wise ℓ_2-norm regularization.

3 Experiment Descriptions and Results

We perform two experiments in this paper, **scene classification** and **scene ranking**. First, we divide our 1973 TRs into 50 chunks of time, the first 25 of which are our training data and the latter 25 of which are our testing data. We learn maps in both directions, fMRI → text and text → fMRI on the training data. Suppose we are predicting text from fMRI. For each time chunk $i \in [1, 25]$ in fMRI space, we predict chunk i in text space using the learned map. Then, we calculate the Pearson correlation of the predicted chunk i with each of the true chunks $j \in [1, 25]$, and we rank the chunk indexes by correlation. For the **scene classification** task, we report "correct" if the true chunk index is ranked among the top 5 chunks produced by this sorting. This measure has a 20% chance rate, and closer to 1 is better. For the **scene ranking** task, we simply report $1 - \frac{\text{average rank of the true index}}{25}$. This measure has 50% chance rate, and closer to 1 is better. We report both of these metrics because the 20% chance rate task gives a better idea of the distribution of the ranking, while other authors have used the 50% chance rate, obtaining ranking scores between $70\% - 80\%$ ([3, 11, 12]).

In Figure 1, we display the top accuracy over all algorithmic choices for our four experiments. The DMN region outperforms the others, supporting previous work by [9] and others demonstrating that the DMN plays an important role in narrative processing. We achieve high accuracy performance, reaching 90% for the scene ranking task for text → fMRI.

We now report the percentage of the four experiment types × five ROIs = twenty experiments which used a certain step in our pipeline. 100% of the top performing algorithmic choices for each of the twenty experiments used SRM as the averaging space, 80% used weighted semantic vector aggregation, 80% used Procrustes over ridge regression, and 65% used temporal zero mean. Restricting to fMRI → text experiments, results in 100% usage of temporal zero mean, underlining its importance for fMRI → text.

Table 1 gives the ratio of the performance between algorithmic options, demonstrating the importance of each step. For entry i, we fix option i and average over all other options. For instance, when comparing SRM to Average, we have $2^3 * 5$ ROIs $= 40$ options. Each proposed pipeline modification improves accuracy on average. Weighting the vectors according to inverse frequency also

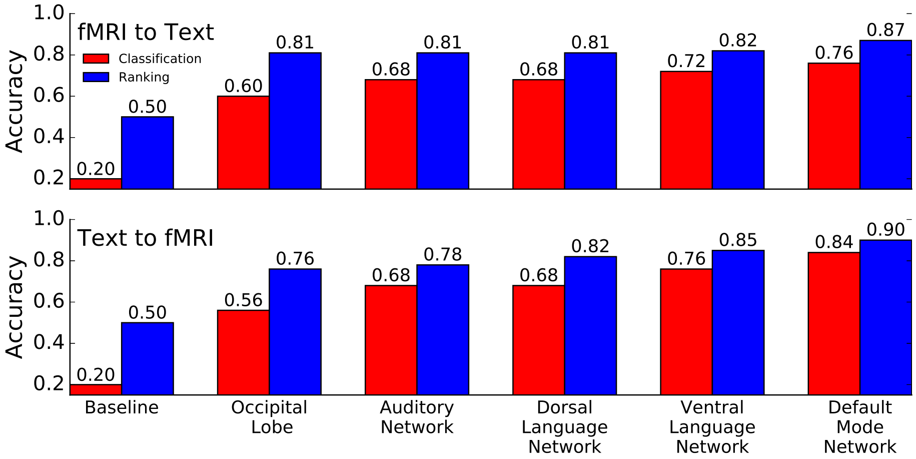

Fig. 1. Best Bidirectional Accuracy Scores for Each Brain Region of Interest for both Scene Classification and Ranking (std. err. over different average subsets < 0.01)

Comparison on the Classification Task	fMRI → Text	Text → fMRI
20-dim SRM / Avg	1.36 ± 0.07	1.16 ± 0.04
Weighted / Unweighted Semantic Vectors	1.16 ± 0.04	1.15 ± 0.015
Temporal Zero Mean / No Zero Mean	2.04 ± 0.08	0.99 ± 0.01
Procrustes / Ridge	1.21 ± 0.07	1.15 ± 0.03

Table 1. Average Improvement Ratio for Various Comparisons

improves upon performance by 1.16× over standard averaging. 20-dimensional SRM improves on average by 1.36× and Procrustes is the regularization of choice with 1.21× improvement in accuracy. Applying temporal zero mean doubles the performance of models mapping fMRI to text. We hypothesize that the temporal mean represents the constant topic of the corpus, which remains relatively unchanging for a Sherlock episode. Since the true semantic vectors are highly correlated before this step, subtracting this invariant allows correlation between predicted and true annotation vectors to distinguish unique time chunks. Interestingly, the performance of the text to fMRI task with this step is left unchanged: Distinguishing fMRI vectors is easier.

Also, the improvements from each algorithmic option are not independent: If we look at the overall improvement ratio for using SRM, weighted averaging, temporal zero mean, and Procrustes versus using none of these options for fMRI to text, we get an overal 1.36× improvement.

In Figure 2, we display the performance of every combination of conditions for the best performing region, the Default Mode Network. We present the results in sorted order to get a sense of how changing conditions affects accuracy. Examining the top five conditions demonstrates that SRM, weighted aggregation, and Procrustes are better for both map directions. More top performers use temporal zero mean for fMRI to text than for text to fMRI. We also see that

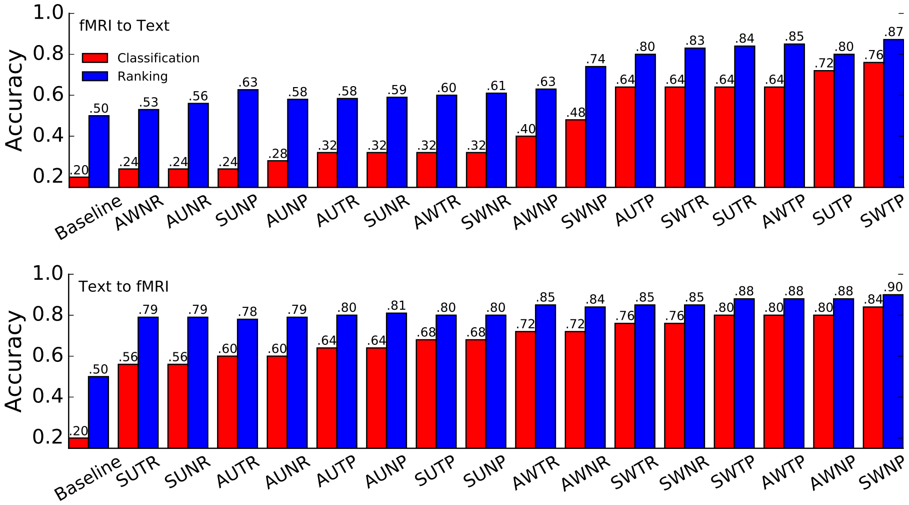

Fig. 2. DMN Bidirectional Accuracy Scores for Scene Classification and Ranking. The acronyms stand for combinations of methods, with the following key: S/A = SRM/Average, W/U = Weighting/No Weighted, T/N = Temporal Zero Mean/No Temporal Zero Mean, P/R = Procrustes/Ridge (std. err. over different average subsets < 0.01)

accuracy has a large range for scene classification $((0.24, 0.76), (0.56, 0.84))$, and thus the algorithmic choices made do matter for our final performance result.

4 Conclusion

The method of combining word vectors is the most essential part of our results.

We demonstrate that different approaches for aggregating individual elements of a word sequence perform noticeably worse than our best result.

Recognizing the existence of a temporal average semantic vector and removing it from our representation is also a crucial aspect of our methods. Since we use only semantic vectors to featurize a movie stimulus dataset, our work provides additional support for the notion that the distributional hypothesis of word meaning may extend to real life multi-sensory stimuli.

Another central result is the improvement of the Shared Response Model over directly averaging the word vectors.

We are able to use multiple subjects to learn a 20-dimensional shared space for the fMRI data which increases performance on our experiments. Thus we provide concrete evidence towards the hypothesis made in [10] regarding the existence of a **shared** fMRI representation across multiple subjects which correlates significantly with **fine-grained** semantic context vectors derived via statistical word co-occurrence properties.

We also demonstrate that linear maps from the brain's Default Mode Network to semantic space achieve better performance on our experiments than maps from other brain regions involved in understanding a movie, including auditory, visual, and language areas.

Bibliography

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

[2] Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T.: Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. PLOS ONE **9** (2014)

[3] Pereira, F., Lou, B., Pritchett, B., Kanwisher, N., Botvinick, M., Fedorenko, E.: Decoding of generic mental representations from functional MRI data using word embeddings. bioRxiv preprint (2016)

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

[5] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). (2014) 1532–1543

[6] Arora, S., Li, Y., Liang, Y., Ma, T., Risteski, A.: A Latent Variable Model Approach to PMI-based Word Embeddings. Transactions of the Association for Computational Linguistics **4** (2016)

[7] Arora, S., Liang, Y., Ma, T.: A principled approach to combine topic models and word embeddings. manuscript (2016)

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

[9] Regev, M., Honey, C.J., Simony, E., Hasson, U.: Selective and Invariant Neural Responses to Spoken and Written Narratives. J Neurosci **33** (2013) 1597815988

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

[11] Pereira, F., Detre, G., Botvinnick, M.: Generating text from functional brain images. Frontier in Human Neuroscience (2011)

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