NeuroImage 180 (2018) 223–231

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Mapping between fMRI responses to movies and their natural language annotations

Kiran Vodrahalli [a,*], Po-Hsuan Chen [a], Yingyu Liang [a], Christopher Baldassano [a], Janice Chen [b], Esther Yong [c], Christopher Honey [b], Uri Hasson [a], Peter Ramadge [a], Kenneth A. Norman [a], Sanjeev Arora [a]

[a] Princeton University, United States
[b] Johns Hopkins University, United States
[c] University of Toronto, United States

ARTICLE INFO

Keywords:
FMRI
Text annotations
Natural language processing
Shared response model
Natural movie stimulus
Multi-modal model

ABSTRACT

Several research groups have shown how to map fMRI responses to the meanings of presented stimuli. This paper presents new methods for doing so when only a natural language annotation is available as the description of the stimulus. We study fMRI data gathered from subjects watching an episode of BBCs Sherlock (Chen et al., 2017), and learn bidirectional mappings between fMRI responses and natural language representations. By leveraging data from multiple subjects watching the same movie, we were able to perform scene classification with 72% accuracy (random guessing would give 4%) and scene ranking with average rank in the top 4% (random guessing would give 50%). The key ingredients underlying this high level of performance are (a) the use of the Shared Response Model (SRM) and its variant SRM-ICA (Chen et al., 2015; Zhang et al., 2016) to aggregate fMRI data from multiple subjects, both of which are shown to be superior to standard PCA in producing low-dimensional representations for the tasks in this paper; (b) a sentence embedding technique adapted from the natural language processing (NLP) literature (Arora et al., 2017) that produces semantic vector representation of the annotations; (c) using previous timestep information in the featurization of the predictor data. These optimizations in how we featurize the fMRI data and text annotations provide a substantial improvement in classification performance, relative to standard approaches.

1. 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 often 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 (Mitchell et al., 2008). The semantic representation of a word was computed using word embeddings, a tool from natural language processing (Deerwester et al., 1990) that represents each word as a point in a *d*-dimensional meaning space. This work was extended (Pereira et al., 2011) to perform "brain reading", using fMRI readings and a popular text-analysis tool called topic modeling to reconstruct word clouds from brain activity evoked by a word/concept stimulus.

The next obvious step in this research program is to understand fMRI readings collected from subjects as they process more complex stimuli such as movies. In such settings, it is not clear how to represent the semantics of the stimulus, since a multitude of signals (auditory as well as visual) are presented within a short time interval. One approach to solving this task was presented in Huth and Nishimoto (2012), which studied fMRI responses to a natural movie stimulus. In this case, the movie stimulus was represented with a feature space of 1705 distinct nouns and verbs. A subsequent study (Huth et al., 2016) examined fMRI responses to audio stories, and departed from the previous work by applying distributional embeddings to featurize the dialog and predict voxel activation. The goal in these papers was to derive a semantic word

* Corresponding author.
E-mail address: kiran.vodrahalli@columbia.edu (K. Vodrahalli).

https://doi.org/10.1016/j.neuroimage.2017.06.042
Accepted 19 June 2017
Available online 23 June 2017

map for the voxels of the brain. Another paper (Wehbe et al., 2014) gathered fMRI data from subjects reading a story, and used unweighted averages of distributional embeddings to featurize sentences for predicting voxel activity.

In this paper, we study the Sherlock fMRI dataset (Chen et al., 2017), which consists of fMRI recordings of 16 people watching the British television program "Sherlock" for 50 min; this time series was broken into 1973 TRs, where each TR corresponds to 1.5 s of film. 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 examine brain data from predefined regions of interest (ROIs) in the brain, and separately analyze each one. In particular, we examine the default mode network (DMN), dorsal and ventral language areas, the occipital lobe, and a 26000-voxel mask containing voxels with high intersubject correlation across the whole brain. We seek to determine whether various modifications to fMRI and text featurization as well as the usage of previous timepoint information help to improve bidirectional mappings between fMRI data and semantic meaning vectors. In particular, we examine the effects of three featurization methods for fMRI and text data: *Low-dimensional shared fMRI representation* across subjects, *weighted semantic embeddings* of text annotations, and using *previous timepoints* in the performance of linear maps between people.

Aggregating fMRI responses across subjects. In prior work, combining fMRI response data from multiple subjects is often solved by averaging, anatomical alignment and smoothing, or latent multivariate feature modeling (Wehbe et al., 2014; Conroy et al., 2013; Huth et al., 2016). Further work concludes that high-level representations of content from movies are shared across people and that there can be considerable de-noising benefits from averaging across people (Chen et al., 2017). Another recent paper (Chen et al., 2015) introduced the Shared Response Model (SRM), an algorithm that stems from previous work on hyperalignment (Haxby et al., 2011). The SRM in Chen et al. (2015) optimizes the objective $\sum_{i=1}^{n} \| X_i - W_i S \|_F$ for a low-dimensional shared space S and orthogonal-column subject specific maps W_i, and can be thought of as a multi-subject extension of PCA. Simultaneously reducing dimensionality across subjects outperforms other averaging approaches at matching up specific timepoints in a movie across subjects.

Semantic representation of stimulus. To find semantic representations of English annotations, it is natural to draw upon related work in natural language processing. One common approach involves word embeddings created by using co-occurrence information in a large corpus like Wikipedia. A simple technique for representing longer pieces of text is to average the vectors for the individual words (Wehbe et al., 2014). Recently, this simplistic idea has been extended in natural language processing by using recurrent neural nets (Kiros et al., 2015) or by modifying the original model for learning word vectors to learn word sequence chunks (for instance, paragraphs) directly from the text (Le and Mikolov, 2014). These more powerful methods have the drawback of requiring large corpora, making them unusable in our current setting where we only have 1973 brief text annotations. Very recently, Arora et al. (2017) suggested a simpler method for this task that requires no additional information beyond the existing word embeddings, yet beats these more complicated methods in standard natural language tasks. We adapt this method to construct *annotation embeddings* using weighted combinations of the vector representations for the words in each annotation. One of our key results is that this new embedding significantly outperforms unweighted averaging of word vectors.

Using previous timestep information. A movie stimulus naturally breaks up into multi-timestep scenes that occur at different timepoints. Thus, at any given timepoint, there may be a window of previous timesteps that are part of the current scene and thus are relevant to understanding the current time point in both fMRI and Text space. We would like to incorporate this past information shared within scenes in order to learn better maps between fMRI and Text. Other models (Huth et al., 2016; Wehbe et al., 2014) incorporate past information by modeling the hemodynamic response function (HRF) that describes the fMRI BOLD response to a stimulus. However, this approach focuses on small time-scales, and only accounts for the delayed and temporally-smeared BOLD response rather than attempting to aggregate scene information. Our approach is to first approximate the HRF delay with a simple one-time shift of 4.5 s, and to then incorporate longer time-scales into our model by including in the featurization a k-sized window of previous timesteps, where k is varied from 0 to 30 (these numbers correspond to 0 to 45 s).

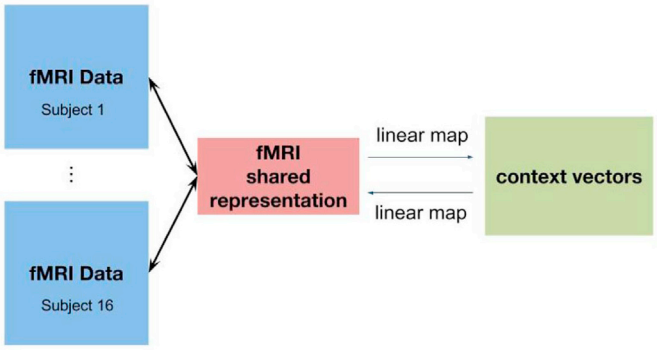

Fig. 1. Summary of Experimental Setup: We learn a shared response for the brain activity of 16 different subjects watching BBC's Sherlock, construct semantic featurizations for associated semantic annotations, and learn bidirectional linear maps between the two data modes.

To evaluate the effect of each of these featurization methods, we use linear maps to relate the fMRI signal to the representation of the semantic content, using only the first half of the movie. These maps are validated with two experiments: scene classification and scene ranking. We divide up the second half of the movie into 25 uniformly-sized chunks. *Scene classification* is the task of using correlation to match predicted intervals of fMRI or semantic activity with the ground truth, and reporting the percentage of the time that the match is perfect. Since there are 25 intervals, random chance performance at this task is 4%. *Scene ranking* is the same task, except we measure the average rank of the correct answer: Random chance performance here is 50%. For a visual summary of the setup, see Fig. 1. These experiments are executed with the fMRI → Text maps (given fMRI data, predict text annotations) as well as the Text → fMRI maps (give text annotations, predict fMRI data).

1.1. Main results

Our goal in this study is to characterize the usefulness of the representation learned by the Shared Response Model, the importance of including previous time steps, and the ideal method for featurizing text information for future predictive tasks. We measure success by evaluating the performance of different featurization approaches on the scene classification and scene ranking tasks.

Our main results are (i) showing that fMRI responses from multiple individuals can be effectively combined using SRM to improve the matching accuracy (1.3× average improvement over our baseline, the average PCA representation) between the fMRI and the text annotation (Table 1, Figs. 6, 7), (ii) demonstrating that a method for combining word vectors into annotation vectors via a suitable weighting (Arora et al., 2017) for averaging word vectors on average improves 1.2 × over unweighted averaging (Table 1, Figs. 6, 7), and (iii) finding that appropriate inclusion of information from previous time steps yields as much as a 5.3× improvement (on average, 1.8 ×) in tasks measuring the performance of mapping from fMRI to Text (see Fig. 6, Dorsal Language ROI). There are diminishing returns after a certain point to including more time steps: The optimal number seems to be around $5 - 8$ previous time steps. For the Text → fMRI task, using previous time steps decreases performance.

We also report the top performances for each task. For the fMRI →

Table 1
Table of Improvement Ratios for Various Algorithmic Parameters: In this table we give the maximum and average improvement ratios for a specific algorithmic technique over another, including usage of previous time steps, SRM/SRM-ICA versus PCA, SIF-weighted annotation embeddings versus unweighted annotation embeddings, and Procrustes versus ridge regression for both fMRI → Text and Text → fMRI. When we use previous timesteps, we consider the results for using 5 − 8 previous time steps. These numbers are all for the scene classification task. Note that the values from the maximum columns can be seen visually in Figs. 6 and 7 respectively.

fMRI → Text	Maximum	Average
Previous Timesteps vs. None	5.3 ×	1.8 ×
Procrustes vs. Ridge	2.8 ×	1.3 ×
SRM/SRM-ICA vs. PCA	1.8 ×	1.3 ×
Weighted-SIF vs. Unweighted	1.6 ×	1.2 ×
Text → fMRI	**Maximum**	**Average**
Previous Timesteps vs. None	2.5 ×	0.5 ×
Procrustes vs. Ridge	3.0 ×	0.8 ×
SRM/SRM-ICA vs. PCA	2.3 ×	1.2 ×
Weighted-SIF vs. Unweighted	1.8 ×	1.1 ×

Text task, our top scene classification performance is 72% accuracy, meaning that for 72% of the time intervals we examine, our predicted annotation representation correlates the most with the true annotation representation for that time interval (see Fig. 5, Whole Brain ROI). Notably, this result improves considerably over the random guessing rate of 4%. The corresponding scene ranking performance is 96%, meaning that on average, the rank of the true annotation representation is within the top 4% when sorted by correlation with the predicted annotation representation. This number implies that, on average, the correct scene is ranked in the top 1 or 2 over all 25 scenes — since there are several more than 1 or 2 scenes in the movie with a single, prominent character like Sherlock or John Watson present, our results imply that our methods can distinguish between scenes that are similar with respect to measures like the presence of certain characters.

On the other hand, the results for Text → fMRI classification were somewhat worse, although performance was still well above chance. The top scene classification performance for Text → fMRI is 56% accuracy (vs. 4% chance), and the corresponding scene ranking accuracy is 91% (vs. 50% chance; see Fig. 5, DMN-A ROI).

2. Methods

2.1. Preprocessing the dataset

The dataset we work with in this paper is Sherlock fMRI dataset (Chen et al., 2017): 16 people watch the "Sherlock" television show for 50 min broken into 1973 TRs, where each TR is 1.5 s of film.

Before performing any analysis, the fMRI data are preprocessed and standardized using the techniques described in Chen et al. (2017). Then, we identify six distinct brain regions of interest (ROIs) that 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. We use the ROIs for the default mode network (DMN-A, DMN-B) and the ROIs for the ventral and dorsal language areas identified in Simony et al. (2016). Our methodology for finding the default mode network relies on intersubject functional correlation (ISFC), a technique first introduced by Hasson et al. (2004). The central idea is that natural stimuli (like movies) evoke reliable, time-dependent activity across a variety of brain networks. For more details, see Fig. 2. We are interested in the DMN ROIs in particular since prior work has demonstrated that these regions play a crucial role in tracking the narrative in settings such as watching movies or reading stories (Hasson et al., 2004, 2010; Honey et al., 2012; Regev et al., 2013; Ames et al., 2015; Simony et al., 2016; Yeshurun et al., 2017). The "Whole Brain" ROI is a 26000-voxel mask of the brain that highlights voxels that have intersubject correlation > 0.2 on the data, and the Occipital Lobe ROI is defined from the MNI Structural Atlas in FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases). We include these ROIs for holistic comparison across the whole brain.

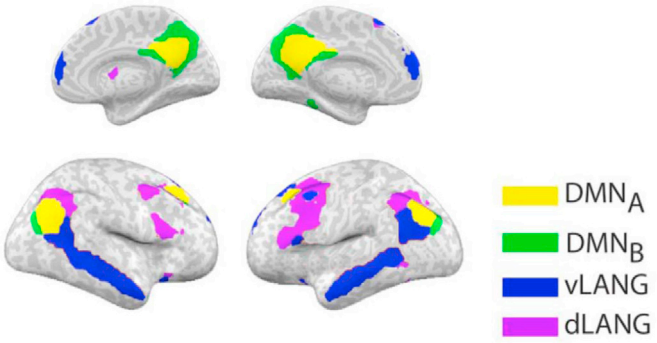

Fig. 2. Visualization of the DMN and Ventral/Dorsal Language Area ROIs (Simony et al., 2016): Here, we display four of the regions of interest on a brain map. These masks were collected on the Pie Man dataset (Simony et al., 2016), then fit to a standard anatomical brain (MNI152), and interpolated to 3-mm isotropic voxels (Simony et al., 2016). In order to define the DMN-A and DMN-B regions, as well as the Ventral and Dorsal language area regions, the intersubject functional correlation matrix (Hasson et al., 2004) was calculated from the fMRI data of 36 subjects collected while they were listening to stories (Simony et al., 2016). Then, *k*-means clustering was applied to find the networks. The DMN-A and DMN-B networks were identified by comparing the resultant clusters to the DMN ROIs derived by thresholding the functional correlation between the posterior cingulate (identified from an atlas) and the rest of the brain in resting-state fMRI data from 36 subjects (Simony et al., 2016). The Ventral and Dorsal language areas were identified by comparing the clusters to previous results in the literature (Simony et al., 2016).

We also truncate the first three TRs of fMRI data and the last three TRs of semantic annotation data. This operation effectively aligns the fMRI and semantic data under the assumption that there is a 4.5 s delay between the onset of the stimulus and the BOLD response signal.

2.2. Constructing and aggregating semantic vectors

The Sherlock fMRI data are supplemented by text annotations describing each TR with a few sentences. These annotations were created by human annotators who viewed the film and carefully noted a couple sentences' worth of detail for every TR. For instance, a moment from the scene where Sherlock and John first meet is described as "Sherlock takes the phone that John hands to him. He flips the screen up, presses a button and blatantly asks: Afghanistan or Iraq?". Notably, these annotations contain some inferences about the subtext of the actions, expressions, and atmosphere of the scene. Adding this information diversifies the information content of the annotations, and allows for the creation of diverse and informative embeddings that capture more semantic content than would otherwise be possible with current natural language understanding techniques. The improved diversity of the embeddings makes the task of identifying unique scenes in text-space easier as well.

In order to represent words, we take advantage of the distributional properties of words in a large corpus – namely, English Wikipedia. We train word embeddings as described in Arora et al. (2016), which perform on par with other standard word embedding techniques like GloVe and Word2Vec (Arora et al., 2016). Now, we diverge from the prior work by calculating and applying a domain specific re-centering of the embeddings. After creating an embedding for each word in the vocabulary of the Sherlock annotations, we calculate the top principal component of all word embeddings in the vocabulary. We then scale the normalized top principal component by the average Euclidean norm of a word embedding in the Sherlock vocabulary. This vector represents a kind of average topic for the Sherlock vocabulary. Since we would like our word embeddings to be discriminative within this average topic, we

algebraically subtract out this component. We can view this step as finding a translation operation that moves the word embeddings away from the region of semantic space that is close to generic words in the Sherlock annotation corpus.

The central assumption in Arora et al. (2016) is the probability model for a word w in a vocabulary V given a context c, where the context represents a small window of words in the corpus. This model is given by $\mathscr{P}[w|c] = \frac{1}{Z_c}\exp(v_w^T c)$ where v_w represents the vector for a given word and Z_c is a term that normalizes the distribution. The idea is that the context vector c represents the subject matter of the text at a given point in time.

Using this assumption and a few others, the word vector learning problem is phrased in Arora et al. (2016) as the *squared-norm objective*:

$$\min_{\{v_w\}_{w\in V}, C} \sum_{w_1,w_2} X_{w_1,w_2}\left(\log(X_{w_1,w_2}) - \| v_{w_1} + v_{w_2}\|_2^2 - C\right)^2$$

where C is a bias term, X is the co-occurrence count matrix between single words in a small window of text (fixed at ≈ 5 words) and v_w are the word vectors we are trying to learn. This objective can be optimized with gradient descent. For a full treatment of the theoretical properties of the word vectors and the derivation of the squared-norm objective, see (Arora et al., 2016).

For every 1.5-second time-point in our Sherlock movie, annotators were asked to provide a natural description of what is happening in the movie: actions, dialog, and so on. This annotation is typically a few sentences long, and contains around 15 words on average. We can think of each annotation as the current context of the movie narrative. The log-linear probability model of Arora et al. (2016) for words given context c implies that the maximum likelihood estimator of the context is simply the average of all words in the annotation. (This formulation is a theoretical justification for a standard rule of thumb in natural language processing for representing the sense of a small piece of text by the average of the embeddings for the words in the text). We will call these representations the **unweighted** annotation vectors.

However, one imagines that not all words in the annotation are equally important, and that a better representation might be possible by taking this idea into account. This approach has been studied in various neural network frameworks (Kiros et al., 2015); however, applying these kinds of models requires a large annotation corpus, while we only have 1973 15-word annotations. A recent paper (Arora et al., 2017) suggests a principled approach for computing a representation of a small piece of text. The intuition from Arora et al. (2017) is that words that 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. Therefore, more frequent words should be weighted less. The paper (Arora et al., 2017) modifies the above language generation model as follows: For a word w given context c, the probability of a word w given context c is

$$\mathscr{P}[w|c] = \alpha\mathscr{P}[w] + (1-\alpha)\frac{\exp(v_w^T c)}{Z_c} \tag{1}$$

where Z_c normalizes the distribution and $\alpha \in [0,1]$. We can think of this model as a weighted sum 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.

The revised 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 $\beta \approx 10^{-4}$. These representations are called the **smooth inverse frequency (SIF)** annotation vectors, or **weighted** annotation vectors. Fig. 3 depicts a example sentence with the respective word weights colored according to importance in the sentence embedding.

Using either the unweighted or weighted approach will produce one annotation vector for each of our T time steps. 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. Here, we assume that the average annotation vector is invariant, which turns out to be a good assumption.

2.3. Shared Response Models for Multi-Subject fMRI

The Shared Response Model (SRM) (Chen et al., 2015) is an unsupervised probabilistic latent variable model for multi-subject fMRI data under a time-synchronized stimulus. From each subject'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 maps W_i with orthogonal columns 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 (Chen et al., 2017). We also experimented with using $k = 50, 80, 100, 1000$, but the results barely varied from using $k = 20$ dimensions. 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.

We also examine a variant of SRM called SRM-ICA (Zhang et al., 2016) that modifies the SRM algorithm with an independent components analysis (ICA) objective. ICA is an unsupervised learning technique that identifies independent signals from a mixture by looking for rotations of the data that produce non-Gaussian signals. SRM-ICA brings this approach to learning a shared space: While in SRM we alternated by solving for W_i by minimizing $\| X_i - W_i S\|_F$ and updating S with the average of $W_i^T X_i$, we change the objective we use to update each W_i to an ICA objective: Maximizing the non-Gaussianity of the shared response $S = \frac{1}{n}\sum_{i=1}^n W_i^+ X_i$, individually with respect to each (X_i, W_i) pair.

Here we are using the Shared Response Model to highlight aspects of the neural signal that are shared across people. To the extent that the information in the text annotations is reflected in the thoughts of most or all subjects, SRM should be helpful in mapping between fMRI and this (shared) information. Note that an alternative use of SRM is to take the shared variance identified by SRM and subtract it out, thereby highlighting idiosyncratic neural variance; this could be useful in situations where fMRI is being mapped to more idiosyncratic cognitive states. In the original Shared Response Model paper (Chen et al., 2015), the authors include an experiment where they apply SRM to two groups that were expected to have differing perceptions of the stimulus. They found that subtracting out shared variance across the groups using SRM improved subsequent discrimination of fMRI time series from the two groups.

In our experiments, we use the implementation of SRM due to Anderson et al. (2016). We compare average SRM and SRM-ICA projections ($\frac{1}{N}\sum_{i=1}^N W_i^T X_i^{\text{test}}$) against the baseline average principal components analysis (PCA) projections. PCA is a standard linear dimensionality reduction technique that finds an optimal (in Frobenius norm) orthogonal projection of the data onto a low-dimensional subspace.

2.4. Learning linear maps

Our approach to predicting semantic annotation 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 SRM, SRM-ICA, or PCA) 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 $X \to Y$, decoding fMRI vectors into semantic space. Our other approach is given by the

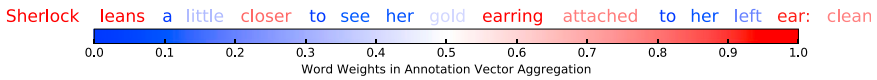

Fig. 3. Visualization of Semantic Annotation Vector Weightings: We display an example sentence from the Sherlock annotations, where we have colored important words red, and unimportant words blue. Brighter red means more important, and darker blue means less important.

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.

2.5. Adding Previous Timesteps

One could augment the fMRI and annotation vectors using past time steps by finding a complicated combination of the features at each time step, resulting in a representation with the same number of dimensions. For now, we sidestep the complexity of this task by simply concatenating k previous vectors to the predictor vector at each time step (TR) before learning mappings as before. A potential downside to this approach is that we linearly increase the dimensionality with k, which can be intractable for large k. However, this approach allows every predictor feature at every timepoint to have its own weight in the linear map, creating a powerful model. Thus, in the fMRI → Text case, we stacked the k previous fMRI vectors onto each fMRI vector, and did not modify the text annotation vectors. In the Text → fMRI case, we stacked k previous text annotation vectors and left the fMRI vectors unmodified. When previous time steps do not exist, we append an all-zeros vector instead. We can think of the modified representations as capturing a notion of the dynamics occurring over an interval of $1.5(k+1)$ (TR length × total number time points) seconds. In this paper, we tried $k = 1 - 9$ in steps of 1, and then $k = 10 - 30$ in steps of 5. See Fig. 4 for a visualization.

2.6. Experiment descriptions

First, we divide our 1973 TRs into 50 uniformly-sized 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 both from fMRI to text annotations and from text annotations to fMRI on the training data. From now on, we refer to **fMRI → Text** experiments as those which take an fMRI representation as input and attempt to predict a semantic annotation vector representation. Likewise, **Text → fMRI** experiments are those which take in a semantic annotation vector input and predict an fMRI representation. Also note that we train the linear maps on the individual TRs as opposed to the 25 chunks.

We perform two primary experiments in this paper, **scene classification** and **scene ranking**. These experiments are applied to both the fMRI → Text and Text → fMRI settings. In the following description, we denote the predictor space by X and the target space by Y.

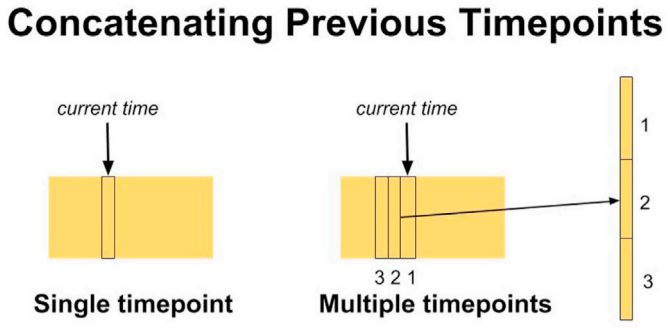

Fig. 4. Visualizing Concatenation: We visualize what the single timestep case looks like compared to a case where we use the previous two timesteps in our featurization as well. The latter case results in a more complicated model, since one of the dimensions of our linear map triples in size.

Suppose we are in the $X \rightarrow Y$ setting. For each time chunk $i \in [1, 25]$ in X-space, we predict chunk i in Y-space using the learned map, by applying the map individually to each TR within the time chunk. Then, we calculate the Pearson correlation of the predicted chunk i (represented by concatenating the representations for each TR in the chunk into one long vector) with each of the actual time chunks $j \in [1, 25]$, and we rank the chunk indexes by correlation.

Scene classification. Given the ranking of actual time chunks by correlation with the predicted chunk, we report the proportion of the time that the correct chunk index is ranked the highest. This measure has a 4% chance rate, meaning that if we randomly ranked the actual chunks, any particular chunk would be the top chunk 4% of the time.

Scene ranking. Given the ranking of actual time chunks by correlation with the predicted chunk, we calculate $1 - \frac{\text{average rank of the correct index}}{25}$. This measure has 50% chance rate, meaning that if we randomly ranked the actual time chunks, the average rank of any particular chunk would be in the middle.

We report both of these metrics because the 4% 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% (Pereira et al., 2011; Wehbe et al., 2014; Pereira et al., 2016).

We also give some additional analysis of the properties of stacking previous time points, and discuss how they affect prediction capabilities. In particular, we observe the dependence of classification accuracy on the number of previous time steps.

3. Results

3.1. Top absolute performances over all algorithms

Fig. 5 suggests that the DMN regions perform well in the experiments, which fits with prior research in this area (Regev et al., 2013; Simony et al., 2016). We achieve 72% accuracy over 4% chance with the Whole Brain region in the scene classification task. Since the scene ranking measure is always ≥ 80%, the average rank of the correct answer is in the top 20% of the scenes, which translates to top 5 scenes out of 25. For fMRI → Text we perform even better, where the average rank of the correct answer is in the top 10% of the scenes (top 3 scenes out of 25). Notably, we get excellent performance out of the Whole Brain region, which has 26000 voxels selected by merely choosing voxels whose intersubject correlation is above a certain threshold. This result demonstrates that our methods are not overly dependent on applying domain-specific knowledge (we do not necessarily have to preselect an ROI to get good results).

fMRI → Text. Here we discuss the performance of the fMRI → Text experiments. In Fig. 5, we display the top accuracy over all algorithmic choices for each experiment. We achieve high accuracy performance, reaching 72% for the scene classification task for fMRI → Text and in the mid-90%s for the scene ranking tasks. In particular, the Whole Brain and the DMN regions perform best, supporting previous work by Regev et al. (2013) and others demonstrating that the DMN plays an important role in narrative processing.

Text → fMRI. On the other hand, we see that the Text → fMRI experiments perform worse than the fMRI → Text experiments. The best top −1 scene classification accuracy performance is 56% for the DMN-A region, and the other top performing regions get accuracy in the mid-to-high 40% accuracy. For the ranking task, performance ranges from 80% − 90%, which is again slightly worse than the fMRI → Text ranking

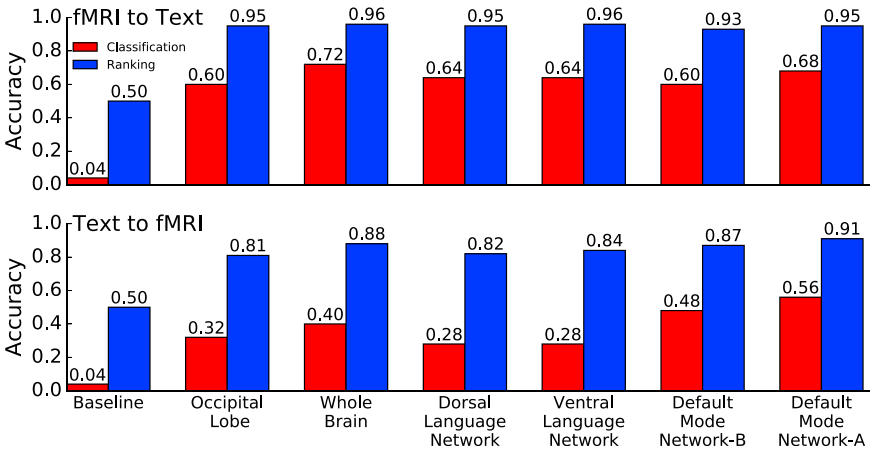

Fig. 5. Best Bidirectional Accuracy Scores for Each Brain Region of Interest for both Scene Classification and Ranking: In this figure, for each ROI and for each experiment (Text → fMRI 4% (red), 50% (blue) chance rates; fMRI → Text 4% (red), 50% (blue) chance rates), we give the best performance as a percentage. For all measures, closer to 100% is better. We can see that Whole Brain, DMN-A, and DMN-B tend to perform the best, and that fMRI → Text performs better than Text → fMRI.

experiment.

3.2. Comparing algorithmic choices

In order to simplify presentation for Figs. 6 and 7, we chose to fix the algorithmic parameters that uniformly outperformed other options. All linear maps for fMRI → Text were learned using the Procrustes method and all linear maps for Text → fMRI were learned using the ridge regression approach. We fixed these for comparison purposes since, for fMRI → Text scene classification, Procrustes performed 1.25 × better than ridge on average (Table 1). On the other hand, ridge performed 1.2 × better than Procrustes on average over Text → fMRI scene classification (Table 1). As a caveat, there were exceptions to the rule, as the max ratios in Table 1 indicate. In Figs. 6 and 7, for the data points that are

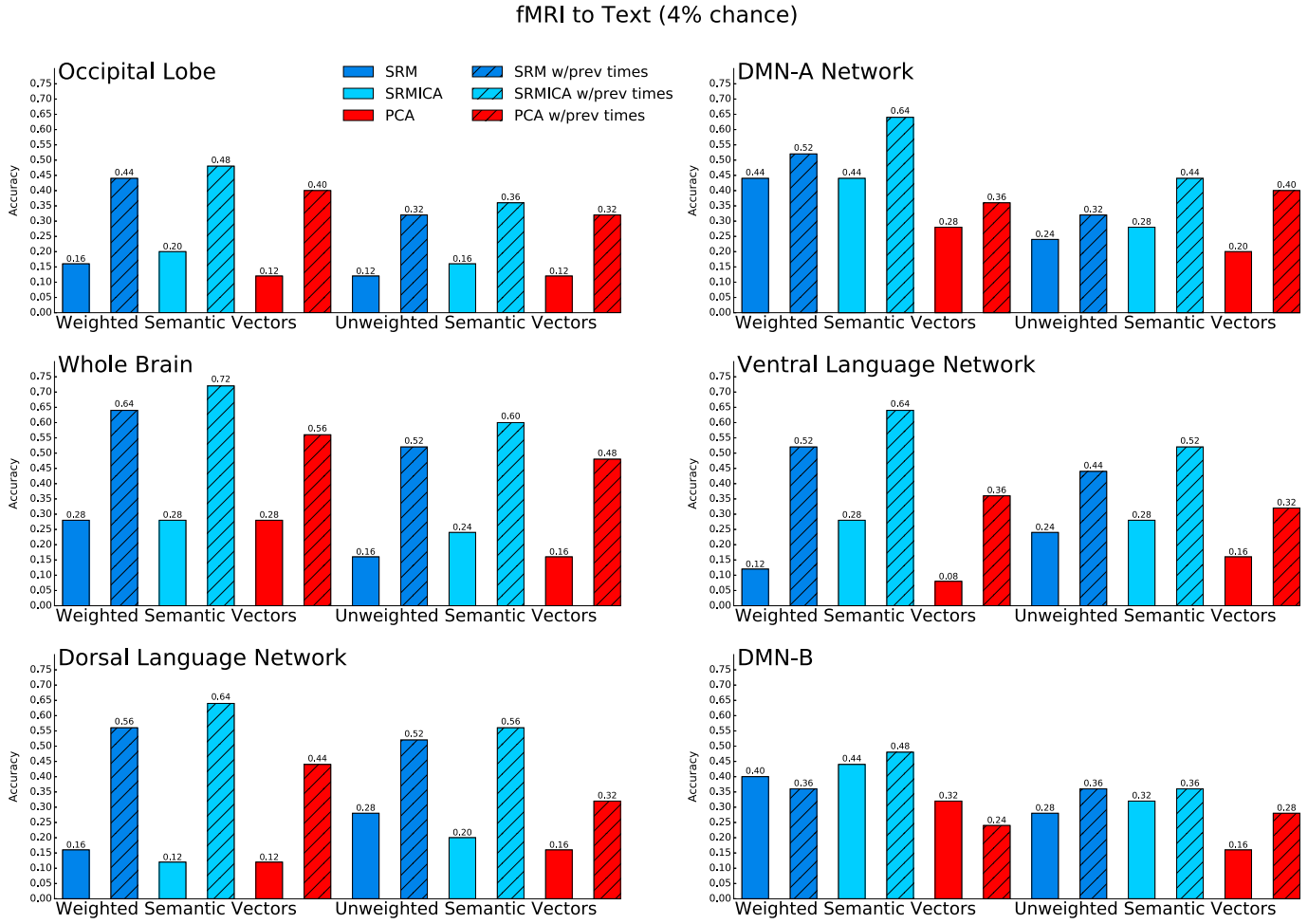

Fig. 6. Comparisons for all ROIs for the fMRI → Text Top-1 Scene Classification Experiment: The chance rate for this task is 4%. Each plot is for a different ROI. Here, we only display results which use the Procrustes linear map since it on average performs better than ridge regression for fMRI → Text. We also fix the number of previous time points used for the shaded bars at 8 previous time steps, since that tends to be near optimal. We present comparisons between SRM/SRM-ICA and PCA using blue colors versus red colors, and compare weighted semantic aggregation (left) to unweighted semantic aggregation (right) by x-axis position.

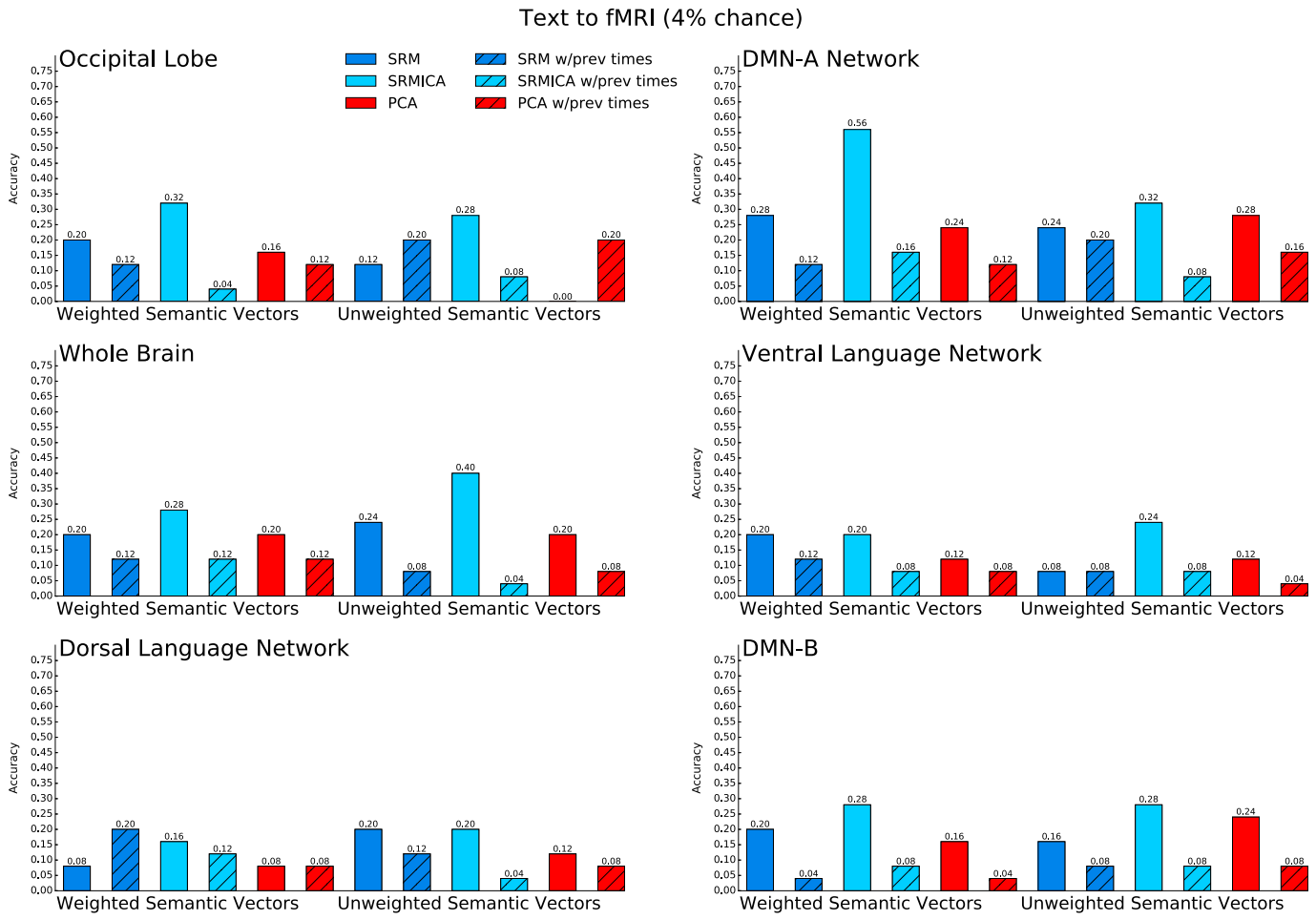

Fig. 7. Comparisons for all ROIs for the Text → fMRI Top-1 Scene Classification Experiment: The chance rate for this task is 4%. Each plot is for a different ROI. Here, we only display results which use the ridge regression linear map since it on average performs better than Procrustes for Text → fMRI. We also fix the number of previous time points used for the shaded bars at 8 previous time steps, since that tends to be near optimal. We present comparisons between SRM/SRM-ICA and PCA using blue colors versus red colors, and compare weighted semantic aggregation (left) to unweighted semantic aggregation (right) by x-axis position.

labeled as using previous time steps, we reported the result for 8 previous time steps. The optimal number of previous time steps for fMRI → Text was typically between $5-8$, and so we fixed that choice of parameter across all of the graphs in these figures.

Comparing SRM and SRM-ICA to PCA. We see considerable improvement on best-case performance when using SRM or SRM-ICA over PCA, particularly on the fMRI → Text tasks, in some cases gaining as much as $1.8\times$ the top -1 scene classification performance of PCA, as demonstrated in Fig. 6. Typically, SRM-ICA tends to perform slightly better, especially on the Whole Brain ROI. The case is weaker for Text → fMRI, since though we can find that performance increases by as much as $2.3\times$ the top -1 scene classification performance, the average benefit is smaller (Table 1, Fig. 7). If we look at average case improvements, we see considerable gains in both directions: SRM/SRM-ICA improve on average by $1.3\times$ over PCA for fMRI → Text scene classification, and on average by $1.2\times$ over PCA on Text → fMRI scene classification. For the ranking tasks, we note that while performance improvement for the best selections of algorithm parameters is not as distinct, SRM and SRM-ICA can drastically improve upon PCA performance for poor selection of parameters. This fact suggests that one should always use SRM or SRM-ICA over PCA, since on new datasets where it is not known which linear map to use, or the number of previous time points to incorporate in the analysis and so on, our results here suggest that these SRM-variants will improve strongly upon PCA if the parameters are poorly chosen, and still improve decently upon PCA otherwise.

Weighted vs. Unweighted Aggregation of Word Embeddings. Using the SIF-weighted embeddings improves upon unweighted averaging when featurizing the annotation vectors as well. Examining Table 1 and Fig. 6, we see that for fMRI → Text top -1, there is improvement on best-case performance by as much as $1.3\times$ by using weighted embeddings. On average, we see that weighted embeddings improve by $1.2\times$ over the unweighted embeddings. Looking at Fig. 7, the case is weaker for Text → fMRI top -1; while for some algorithms and ROIs we see as much as $2.5\times$ improvement on best-case performance by weighted aggregation embeddings, we also see that sometimes unweighted averaging can outperform weighted averaging. However, on average, weighted embeddings improve by $1.1\times$ over unweighted averaged embeddings.

The Effects of Previous Time Points. Fig. 6 demonstrates the positive effect of adding previous time steps to the accuracy scores for the fMRI → Text case. Table 1 demonstrates that at best, using previous timepoints can improve performance by as much as $5.3\times$. On average, this improvement is $1.8\times$, nearly doubling performance. On the other hand, Fig. 7 shows that for Text → fMRI, adding previous time steps almost universally hurts performance and on average halves performance (Table 1). This fact is also evident from Fig. 8, which illustrates the situation for the DMN-A ROI.

Notably, the effect of using previous time steps is different from learning a hemodynamic response function, which other authors (Wehbe et al., 2014; Huth et al., 2016) have done in the past. Instead, we are investigating whether information from longer time scales helps improve performance. In Fig. 8, we see that there are some peaks in classification performance between 5 and 8 previous time steps ago (or $7.5-9.0$

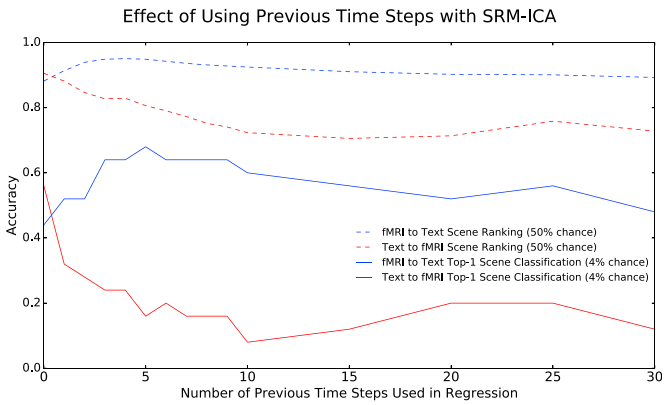

Fig. 8. Varying Previous Timesteps: For the DMN-A region, choosing SRM-ICA, weighted average, Procrustes for the fMRI → Text linear map, and ridge for the Text → fMRI linear map, we plot the relationship between accuracy (y-axis) and number of previous time points used in the linear map fit (x-axis). We can see a peak at around using $5-8$ previous TRs as optimal for the fMRI → Text tasks, and a relatively monotone decay for using any previous TRs in the Text → fMRI tasks.

seconds ago, after having taken into account the HRF). However, using any number of previous time steps (up to as long as 30 TRs ago, or 45 s) still improves over the baseline of using no previous time steps.

For Text → fMRI however, the story is different. We see no improvement in performance when using previous time points, and in fact performance decreases (Fig. 8). We discuss differences between the Text → fMRI and fMRI → Text results in the next section.

4. Discussion

In this paper, we have explored several methods that improve our success at mapping between fMRI response to a natural stimulus and semantic text data describing this stimulus. We see that SRM and SRM-ICA perform considerably better than simple averaging or using PCA. Fig. 6 demonstrates that weighted aggregation of the words in semantic space to form annotation vectors improves the baseline accuracy by a reasonable amount, relative to simple averaging. We also show that adding previous time steps improves accuracy substantially.

Using SRM-ICA in fMRI space, weighted annotation vectors in semantic space, and a Procrustes linear map learned between the concatenations of five previous time points in fMRI and semantic space, we are able to achieve 72% scene classification accuracy over 4% chance rate for the Whole Brain region on the fMRI → Text task.

Other ROIs are typically above 60% scene classification accuracy as well. Similarly, in the scene ranking task, we achieve $>90\%$ average rank for the correct answer across ROIs. Text → fMRI does not perform as well but is still far above chance (56% with DMN-A ROI for 4% chance rate, and $>80\%$ average rank across ROIs). Another takeaway is that SRM and SRM-ICA improve upon PCA almost always, and provide particularly substantial improvement in cases where the other parameter settings (like the semantic featurization or selection of linear map and associated hyper-parameters) are not necessarily tuned. These results indicate that we are able to use multiple subjects to learn a 20-dimensional shared space for the fMRI data that increases performance on our experiments. Thus, we provide concrete evidence towards the hypothesis made in Huth et al. (2016) regarding the existence of a **shared** fMRI representation across multiple subjects that correlates significantly with **fine-grained** semantic context vectors derived via statistical word co-occurrence properties.

The method of combining word vectors is another essential part of our results. We demonstrate that weighted-SIF averaging (Arora et al., 2017) for aggregating individual elements of a word sequence performs on average $1.2\times$ better than unweighted averaging for fMRI → Text top -1 scene classification, and on average $1.1\times$ better for Text → fMRI top -1 scene classification. 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.

Finally, we note that using multiple previous timepoints when mapping from fMRI → Text is very beneficial and significantly improves results by a factor of as much as $5.3\times$, and on average nearly doubles performance (Table 1).

Overall, accuracy for Text → fMRI was worse than for fMRI → Text. Also, using previous time points hurt performance for Text → fMRI (whereas it helped for fMRI → Text). One possible cause of these differences is that, in several places in the movie, text vectors were almost identical between adjacent TRs, whereas the fMRI patterns varied. Where this property occurred, Text → fMRI posed a (more difficult) one-to-many mapping problem, whereas fMRI → Text was an (easier) many-to-one mapping. These "pockets of stationarity" in the text vectors are a consequence of our using human-generated annotations to construct our semantic embeddings. As we noted before, these annotations reflect the annotator's inferences about what is happening in the movie; if the annotator's understanding of the current situation in the movie stays relatively stationary, the text embeddings will also stay relatively stationary. This reasoning may also account for why adding previous timepoints was not helpful for Text → fMRI; if the annotations for previous timepoints are the same as for the current timepoint, adding these previous timepoints has the effect of adding extra free parameters to the model without adding new, useful information.

Putting all of these ideas together, we think it is possible that the annotations left out some details that were nonetheless cognitively registered by the fMRI participant (and thus registered in their fMRI data). This suggests that one way to improve Text → fMRI accuracy would be to identify points in time when text annotations are relatively stationary, and then go back and encourage the annotator to include new details that distinguish between the time points that they may not have noted on the first pass. At the same time, we should note that the human cognitive system's ability to maintain a stable understanding of a situation in the face of changing sensory input is a feature and not a bug; the relative stationarity of the annotations within scenes reflects our ability to extract deep structure from complex narratives, and in ongoing, related work we are developing new tools to identify and study brain regions that are involved in extracting this deep structure (Baldassano et al., 2016).

Acknowledgments

The dataset is online (Chen et al., 2017) and the code used in this paper is available on GitHub (https://github.com/kiranvodrahalli/fMRI_Text_maps_NI). 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, and NSERC Discovery Grant RGPIN 2014-04465 awarded to C. Honey. P.-H. Chen was supported by a Google PhD Fellowship. We also thank C. Chen and V. Mocz for their comments on this work.

References

Ames, D., Honey, C.J., Chow, M., Todorov, A., Hasson, U., 2015. Contextual alignment of cognitive and neural dynamics. J. Cogn. Neurosci. 27, 655–664.

Anderson, M.J., Capota, M., Turek, J.S., Zhu, X., Willke, T.L., Wang, Y., Chen, P.-H., Manning, J.R., Ramadge, P.J., Norman, K.A., 2016. Enabling factor analysis on thousand-subject neuroimaging datasets URL http://arXiv:1608.04647arXiv:1608.04647.

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

Arora, S., Liang, Y., Ma, T. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings, International Conference on Learning Representations (ICLR) URL ⟨https://openreview.net/pdf?Id=SyK00v5xx⟩.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2016. Discovering event structure in continuous narrative perception and memory, bioRxiv preprint URL ⟨http://biorxiv.org/content/early/2016/10/14/081018⟩.

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. Nat. Neurosci. 20, 115–125 (URL ⟨http://www.nature.com/neuro/journal/v20/n1/full/nn.4450.html⟩).

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

Conroy, B., Singer, B., Guntupalli, J., Ramadge, P., Haxby, J., 2013. Inter-subject alignment of human cortical anatomy using functional connectivity. NeuroImage 81, 400–411.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640 (URL ⟨http://media.wix.com/ugd/b75639_74b4709ef98248a1be41e5ea433fdaed.pdf⟩).

Hasson, U., Malach, R., Heeger, D., 2010. Reliability of cortical activity during natural stimulation. Trends Cogn. Sci. 14, 40–48.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representation space in human ventral temporal cortex. Neuron 72, 404–416 (URL ⟨http://haxbylab.dartmouth.edu/publications/HGC+11.pdf⟩).

Honey, C.J., Thompson, C.R., Lerner, Y., Hasson, U., 2012. Not lost in translation: neural responses shared across languages. J. Neurosci. 32, 15277–15283.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224.

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.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S., 2015. Skip-Thought Vectors, Advances in Neural Information Processing Systems URL ⟨https://papers.nips.cc/paper/5950-skip-thought-vectors⟩.

Le, Q., Mikolov, T., 2014. Distributed Representations of Sentences and Documents, Proceedings of the 31st International Conference on Machine Learning, JMLR 32.

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.

Pereira, F., Detre, G., Botvinick, M., 2011. Generating text from functional brain images. Front. Human. Neurosci. 5, 72. https://doi.org/10.3389/fnhum.2011.00072 (URL ⟨http://journal.frontiersin.org/article/10.3389/fnhum.2011.00072⟩).

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

Regev, M., Honey, C.J., Simony, E., Hasson, U., 2013. Selective and invariant neural responses to spoken and written narratives. J. Neurosci. 33, 15978–15988.

Simony, E., Honey, C.J., Chen, J., Lositsky, O., Yeshurun, Y., Hasson, U., 2016. History dependent dynamical reconfiguration of the default mode network during narrative comprehension, Nature Communications 7 URL ⟨http://www.nature.com/articles/ncomms12141⟩.

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

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C.J., Hasson, U., 2017. Same story, different story: the neural representation of interpretive frameworks, Psychological Science URL ⟨https://doi.org/10.1177/0956797616682029⟩.

Zhang, H., Chen, P.-H., Chen, J., Zhu, X., Turek, J.S., Willke, T.L., Hasson, U., Ramadge, P.J., 2016. A searchlight factor model approach for locating shared information in multi-subject fmri analysis, arXiv preprint URL http://arXiv:1609.09432arXiv:1609.09432.