A Semantic Shared Response Model

Abstract

We present the finding that it is possible to identify a semantically-relevant shared representation of fMRI response in an unsupervised fashion using views of multiple subjects watching the same natural movie stimulus. Using the shared response instead of individual subject responses we are able to significantly improve the prediction of voxel values from semantic word vectors representing descriptions of an audio-visual movie.

1. Introduction

Several researchers have attempted to find relationships between word featurizations and fMRI activation in the brain. One popular method due to (Mitchell et al., 2008) gathers fMRI data across several subjects corresponding to text stimuli: individual nouns (Mitchell et al., 2008), a set of words (Pereira et al., 2011), or even a story (Wehbe et al., 2014). Here we address the multi-view nature of finding meaning in the brain. Our specific goal is to determine if an fMRI shared space can be learned across subjects that correlates well with semantic word embeddings. We study the **Sherlock** fMRI dataset (Chen et al., 2016). This consists of fMRI recordings of 17 people watching the British television program "Sherlock" for 45 minutes. In addition, we use externally annotated, sub-second-resolution, English text scene annotations of the program. Using these annotations and the Wikipedia corpus, we employ unsupervised methods to construct semantic context vectors using global co-occurrence matrix factorization and sparse coding (Pennington et al., 2014; Arora et al., 2015; 2016). We then use the unsupervised Shared Response Model (SRM) (Chen et al., 2015) to construct a shared embedding space across the 17 subjects for six distinct brain regions of interest (ROI). Finally, we construct a map from semantic embedding space to the fMRI shared embedding space of our dataset. The models are validated with three experiments: fMRI scene classification, assessing context vector quality, and fMRI reconstruction.

The present work is similar in some ways to that of (Huth et al., 2016), which also sought to map text embeddings from narrative stimuli to fMRI data. The main difference is our use of the Shared Response Model. An additional difference is that we analyze fMRI responses to an audio-visual movie with annotations describing unvoiced aspects of the scenes. In contrast, (Huth et al., 2016) analyzed fMRI responses to auditory narratives for which the spoken text corresponds identically with the word embedding representations.

We provide concrete evidence towards the hypothesis made in (Huth et al., 2016) 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 approaches. Our use of multiple subject views of the movie data plays a great role in boosting the performance of our model and suggests that if the model in (Huth et al., 2016) was applied using multiple-subject SRM, their results would also improve. 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.

2. Semantic Shared Response Model

Our model has three components: a feature space learned from the multi-subject fMRI, a semantic movie context featurization procedure based on natural language processing methods, and a learned mapping from semantic context to fMRI feature space. This permits the use of all subjects' data and creates a bridge from fMRI space to word space.

The Shared Response Model (SRM) (Chen et al., 2015) 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 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. Weighted singular value decomposition then yields low-rank semantic vectors whose geometry clusters similar words and creates linear algebraic analogy relationships (Arora et al., 2015). Recent work has applied sparse coding to these word vectors to get fine-grained 300-dimensional

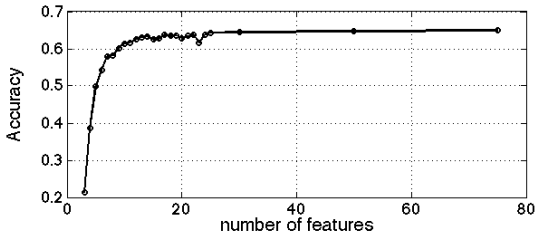

Figure 1. Exp. 1: Scene prediction experiment with SRM

representations of specific word senses (Arora et al., 2016).

To align the fMRI and context vectors, we apply ridge regression to learn a linear map from the context vectors to the shared fMRI space. We then predict $\hat{S}$ from the context vectors and compute the correlation $\langle S, \hat{S} \rangle$. For comparison, we try ridge regression from the context vectors to the individual fMRI responses X_i and compute correlation $\langle X_i, \hat{X}_i \rangle$. Significantly lower testing reconstruction error in the first mapping implies that the distributed context embeddings capture some extrinsic notion of meaning that extends beyond a corpus into real-world stimuli.

3. Experiments

We conduct three experiments to verify the effectiveness of the model: an experiment on fMRI data, an experiment verifying the quality of the context vectors and a final experiment in which we describe the scene using both multi-subject fMRI data and context vectors.

Exp. 1: SRM scene prediction. The movie was divided into 50 scenes by human raters. Here, we learn subject specific mappings from the fMRI data to a shared representation using SRM (Chen et al., 2015), and we verify that the average scene response from a left out subject can be well classified using other subjects' scene responses. To accomplish this, we fit SRM to half of the movie response and learn subject specific transformations $W_i \; \forall i$. These transformations are applied to the average scene response from all subjects. The results (Fig. 1) show the averaged prediction accuracy over 10 random splits on movie scenes and 40 random left out subjects for each split. Chance accuracy level is $1/25$ (25 scenes in each half of the movie). The results suggest that using 20 dimensions captures scene-specific aspects of the fMRI signal, yielding well above chance accuracy (65% versus 4%).

Exp. 2: Evaluating Context Vector Quality After generating 300-dimensional context vectors for each time point in the movie as per the approach in Sec. 2, we check the quality of the vectors by finding nearby vectors of fine-grained meaning, which result from the sparse coding step (Arora et al., 2016). For instance, consider an example annotation of a scene in Sherlock: "Donovan looks up at the reporters and continues: 'Preliminary investigations...' Lestrade looks distressed. Donovan continues: '... suggest that this was suicide. We can confirm that this...". Nearby word vectors correspond to words like "investigation" (corr. $= 0.78$), "suicide" (corr. $= 0.74$), "CNN" and "Reuters" (corr. $= 0.71$), and "police" (corr. $= 0.70$). The other context vectors heuristically have similar quality to this example.

ROIs (Simony et al., 2016)	20-dim SRM	raw fMRI
Ventral Language Network	0.15	0.06
DMN Network	0.11	0.04
Auditory Network	0.11	0.05
Dorsal Language Network	0.10	0.03
Occipital Lobe	0.08	0.04
Early Visual Cortex	0.08	0.04

Table 1. Exp. 3: Comparing corr($\hat{S}, S$) and avg. corr($\hat{X}_i, X_i$)

Exp. 3: Mapping Between fMRI and Context Vectors We now assess the testing performance of the ridge regression described in §2. In Table 1, for each of the ROIs outlined in (Simony et al., 2016), we give the Pearson correlation between the reconstructed voxel-space ($\hat{X}_i$) or shared-fMRI-space ($\hat{S}$) and the true space (X_i, S respectively). The SRM outperforms the individual subject maps substantially. Most interestingly, the ventral language area has the best performance overall.

4. Discussion

We have demonstrated that the multi-view SRM model produces a semantically relevant 20-dimensional space using views of multiple subjects watching *Sherlock*. This low-dimensional fMRI shared space is able to match fMRI responses to scenes with performance considerably above chance. We were also able to construct a 300-dimensional embedding of the semantic context induced by scene annotations. Finally, we bridge fMRI response and scene annotations through a linear transformation. We showed that bridging scene annotations with the fMRI shared space leads to significantly larger correlation than bridging scene annotations with the individual fMRI response. This implies that the shared fMRI space does a better job of highlighting the semantic "signal" compared to individual fMRI views.

For future work, we would like to include an additional visual view from the movie stimulus in an end-to-end architecture incorporating the shared fMRI and semantic embedding views. We are currently developing an autoencoder variant of SRM to be included as a component of such a model. Preliminary tests suggest that this new model performs well, but further testing remains to be completed.

References

Arora, Sanjeev, Li, Yuanzhi, Liang, Yingyu, Ma, Tengyu, and Risteski, Andrej. RAND-WALK: A latent variable model approach to word embeddings. *arXiv:1502.03520*, 2015.

Arora, Sanjeev, Li, Yuanzhi, Liang, Yingyu, Ma, Tengyu, and Risteski, Andrej. Linear Algebraic Structure of Word Senses, with Applications to Polysemy. 2016.

Chen, Janice, Leong, Yuan Chang, Norman, Kenneth A., and Hasson, Uri. Shared experience, shared memory: a common structure for brain activity during naturalistic recall. *bioRxiv preprint*, 2016.

Chen, Po-Hsuan, Chen, Janice, Yeshurun, Yaara, Hasson, Uri, Haxby, James V., and Ramadge, Peter J. A Reduced-Dimension fMRI Shared Response Model. *The 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

Huth, Alexander G., deHeer, Wendy A., Griffiths, Thomas L., Theunissen, Frédérick E., and Gallant, Jack L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458, 2016.

Mitchell, Tom M., Shinkareva, Svetlana V., Carlson, Andrew, Chang, Kai-Min, Malave, Vicente L., Mason, Robert A., and Just, Marcel Adam. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320:1191–1194, 2008.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Pereira, Francisco, Detre, Greg, and Botvinnick, Matthew. Generating text from functional brain images. *Frontier in Human Neuroscience*, 2011.

Simony, Erez, Honey, Christopher J., Chen, Janice, Lositsky, O., Yeshurun, Y., and Hasson, Uri. History dependent dynamical reconfiguration of the default mode network during narrative comprehension. *(in review)*, 2016.

Wehbe, Leila, Murphy, Brian, Talukdar, Partha, Fyshe, Alona, Ramdas, Aaditya, and Mitchell, Tom. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE*, 9, 2014.