CS230

A computational knowledge engine for human neuroscience

Ellie Beam (ebeam@stanford.edu)
Department of Psychiatry and Behavioral Sciences, Stanford University

Abstract

Most mental functions were defined decades ago in psychology before it could be known how they related to brain activity. Here, deep learning is applied to assess how strongly mental functions in human neuroimaging article texts predict spatial locations of brain activity. Several sets of semantic features are compared including term occurrences, embeddings, and context encodings of the titles and full texts. The results support averaging the embeddings of mental function terms in the full text as input for this classification task. Among the top-performing classifiers, feature maps for each brain region lend insight into which mental functions most strongly predict activation across the literature.

1 Introduction

Human neuroimaging seeks to understand how mental functions are represented in activity patterns of the brain. The flow of inquiry, however, has been largely unidirectional -– taking mental functions defined decades earlier in psychology as the premise for brain mapping efforts. Are the mental functions that have been reified in the field actually predictive of brain activation? With 25 years of neuroimaging articles in hand, we will take the mental functions that studies discuss in their texts as features for classifiers predicting where neural activity is reported. Deep learning is applied both to encode article texts and to predict locations of brain activity. The features of the classification models are then used to map out a "brain dictionary" of the terms for mental functions that most strongly predict activity in each neuroanatomical structure.

Before mental functions are given as features to classifiers predicting neural activity, they must be encoded with a language model. Our text inputs include the titles and full texts of nearly 20,000 fMRI and PET studies that reported coordinates in the human brain. In addition to assessing the predictive value of mental functions, a goal of this project is to assess which language model for mental functions offers superior predictive performance. It is a difficult task because neuroimaging articles range in length and may discuss topics not directly related to the neural data they report. These challenges are addressed by using domain knowledge to narrow the feature space (i.e., by using a lexicon of terms for mental functions), and by reducing the dimensionality of the input words with models for their semantic content and context.

2 Related Work

Approaches were previously developed for automated synthesis of human neuroimaging articles. The Neurosynth platform enables automated mapping of associations between brain coordinate data and terms in article texts [1]. The current project will build on associative meta-analysis with a supervised learning procedure. This will involve mapping brain coordinates onto atlas labels and encoding texts into features that are best predictive of them.

CS230: Deep Learning, Spring 2019, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

The simplest language model in this setting represents mental functions by the binary occurrences of terms in article full texts. A disadvantage of using term occurrences as features is that they are high-dimensional and sparse (i.e., there are many terms, and few of those terms occur in a given article), making classification models prone to overfitting. Word embeddings, by contrast, quantify the semantic features of a given word in a dense, lower-dimensional space. Recent work has shown that word embeddings can successfully predict brain activation patterns recorded while individuals listened to natural speech in the fMRI scanner [2]. The global vectors (GloVE) algorithm has outcompeted other word embedding approaches in classification on the document level [3], making it an appropriate choice for our application. Additional improvements in predictive performance have been achieved using long short-term memory (LSTM) encodings [4], which are able to capture semantic content over longer spans of input text. The features for our classification model will thus range from simple term occurrences to GloVe embeddings and LSTM encodings.

3 Dataset and Features

3.1 Lexicon of Mental Functions

The set of possible mental functions for term occurrence and embedding models was defined by a lexicon of 1,542 words and phrases (Figure 1). Terms were compiled from the BrainMap Taxonomy [3], the Cognitive Atlas [4], the Cognitive Paradigm Ontology [5], the Mental Functioning Ontology [6], and the Neuroscience Informatics Framework [7]. The included terms may refer to mental constructs (e.g., "emotional memory"), processes (e.g., "retrieval"), percepts and stimuli (e.g., "face"), or task paradigms (e.g., "face identification task"). Terms were required to occur 5 or more times across the corpora of article full texts.

Lexicon Source	Entities
BrainMap [5]	○ Behavioral Domains ○ Paradigm Classes
Cognitive Atlas [6]	○ Behaviors ○ Concepts ○ Personality Traits ○ Tasks
Cognitive Paradigm Ontology [7]	○ Stimulus Modality ○ Explicit Stimulus ○ Stimulus Role ○ Response Modality ○ Overt Response ○ Instructions ○ Paradigm Classes
Mental Functioning Ontology [8]	○ Mental Functions
Neuroscience Informatics Framework [9]	○ Functions

Figure 1: Sources compiled into a lexicon of mental functions.

3.2 Corpora of Human Neuroimaging Articles

First, a corpus of 18,155 human neuroimaging articles was curated for models predicting brain activation coordinate data from article full texts. Articles were gathered from BrainMap [3] (n = 3,346), then Neurosynth [1] (n = 12,676), and finally by deploying the Automated Coordinate Extractor (ACE) [10] (n = 2,133). The key inclusion criterion was that human brain coordinates were reported in standard Montreal Neurological Institute (MNI) or Talairach space. Second, an expanded corpus of 29,828 human neuroimaging articles was curated to generate GloVe embeddings. This corpus included an additional 11,673 articles returned by the following PubMed search: *((positron emission tomography) OR "fmri") AND (psychiat* OR psychol* OR mental OR emotion* OR reward* OR cognit* OR social* OR arous*) NOT (cancer OR tumor OR stroke OR hematoma OR hemorrhage OR aneurysm OR encephalitis OR infection)*.

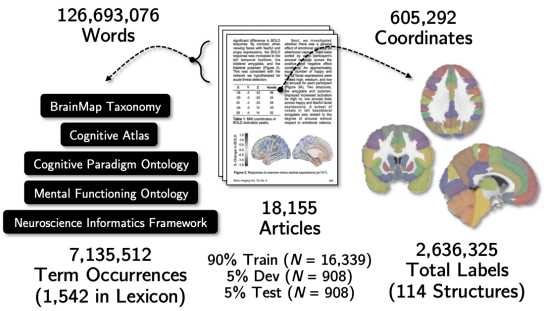


Figure 2: Processing of neuroimaging article texts and coordinate data.

3.3 Natural Language Processing of Article Texts

Article full texts were extracted from PDF or HTML files downloaded through Stanford University subscription services, from PDF files available through the PMC Open Access Subset, or from XML files in the PMC Author Manuscript Collection. Preprocessing of texts and the lexicon included case-folding, removal of stop words and punctuation, and lemmatization with WordNet. *N*-grams listed in the lexicon were combined with underscores in article texts (Figure 2).

3.4 Mapping of Activation Coordinate Data

Brain activation coordinates correspond to spatial locations in the human brain that were found to be statistically related to measures of mental function. A total of 605,292 spatial coordinates were compiled from the BrainMap, Neurosynth, and ACE databases. Coordinates reported in Talairach space were converted to MNI space by the Lancaster transform [11]. MNI coordinates were then mapped probabilistically in a one-to-many fashion onto 114 gray matter structures in a composite neuroanatomical atlas spanning the human cerebrum [12] and cerebellum [13] (Figure 2). This procedure resulted in 2,636,325 total labels for brain structures across the corpus of 18,155 articles.

3.5 Data Splits

The corpus of 18,155 articles was randomly split into sets for training (90%, n = 16,339), dev (5%, n = 908), and out-of-sample testing (5%, n = 908). Training of GloVe embeddings was performed on these texts in addition to the 11,673 texts returned by the PubMed search described above (n = 29,828 total). LSTM and GloVe models trained on full texts (not titles) used a subset of 5,000 training set articles in order to more efficiently search the hyperparameter and model architecture spaces.

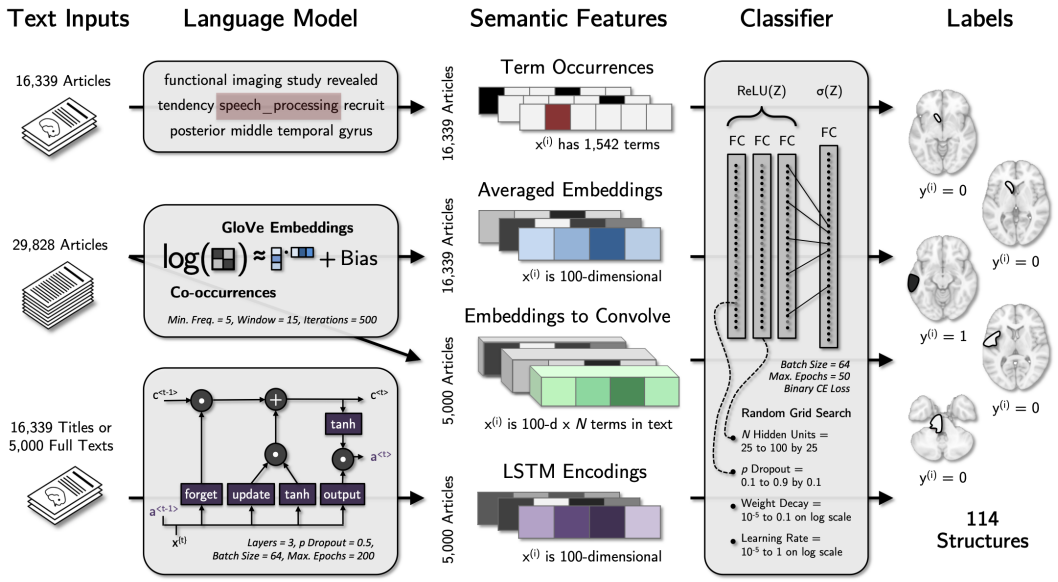


Figure 3: Pipeline for predicting locations of brain activity from neuroimaging article texts.

4 Methods

4.1 Language Models

4.1.1 Term Occurrences

Term occurrence features were generated by computing a document-term matrix with terms for mental functions (Figure 3, top). The matrix spanned the 18,155 articles with coordinate data and the 1,542 terms in the lexicon. Entries were "1" if a term occurred in a given article and "0" if it did not.

4.1.2 GloVe Embeddings

GloVe was fitted on the expanded corpus of 29,828 documents (Figure 3, middle). The model had an embedding dimension of 100 and window size of 15 words over 500 iterations. The vocabulary comprised 350,543 terms that occurred 5 or more times. The code for training GloVe was adapted from `github.com/stanfordnlp/GloVe`.

Document-level features were generated from the GloVe embeddings by two approaches: (1) by averaging embeddings of terms that occurred in each article, and (2) by inputting the embeddings

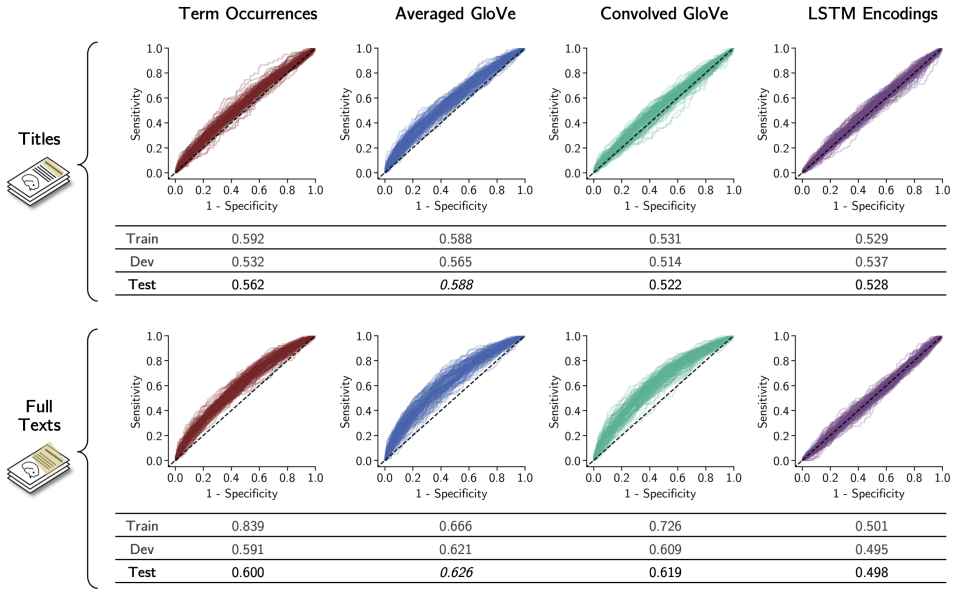
of terms that occurred to a convolutional layer at the head of the neural network. The motivation for the convolutional layer was to enable the network to reduce the dimensionality of the term embeddings without requiring each dimension to have equal weight, as the averaging operation does. The convolutional layer had a kernal equal in size to the embedding dimension by the vocabulary dimension, and its output size matched that of the hidden dimension carried throughout the remainder of the neural network.

4.1.3 LSTM Encodings

Finally, LSTM models were trained to predict word occurrences and subsequently to encode the text inputs (Figure 3, bottom). The LSTM model trained on the full texts was limited to 5,000 documents as noted above (batch size = 16); the model trained on titles took all 16,339 articles in the training set as inputs (batch size = 32). Each model included 3 layers with 100 neurons per layer and a dropout probability of 0.5. As for GloVe, the window size was 15 words. Training was allowed to proceed until no further improvements in loss were observed, which occurred after a single epoch for the model trained on full texts and after 60 epochs for the model trained on titles. The LSTM code was adapted from `github.com/pytorch/examples/tree/master/word_language_model`. Document-level features were generated by forward propagating each article's terms, resulting in a 100-dimensional encoding at the last hidden layer.

4.2 Classification Model

Features from each language model were taken as inputs to a one-versus-all neural network classifier predicting whether an activation coordinate was reported in each of 114 brain structures (Figure 3). The classifier consisted of 3 hidden layers that were fully connected (FC), with the first layer adapted to the input size of features. Hidden layers were activated by rectified linear units (ReLU), while the last layer had a sigmoid activation. Classifiers were trained with a batch size of 64 over 50 epochs. Hyperparameters were tuned on the dev set through a randomized grid search over 50 combinations of the following: units per hidden layer = 25 to 100 by 25, dropout probability = 0.1 to 0.9 by 0.1, L2 weight decay = 10^{-5} to 0.1 on a log scale, and learning rate = 10^{-5} to 1 on a log scale.



	Term Occurrences	Averaged GloVe	Convolved GloVe	LSTM Encodings
Train	0.592	0.588	0.531	0.529
Dev	0.532	0.565	0.514	0.537
Test	0.562	*0.588*	0.522	0.528

	Term Occurrences	Averaged GloVe	Convolved GloVe	LSTM Encodings
Train	0.839	0.666	0.726	0.501
Dev	0.591	0.621	0.609	0.495
Test	0.600	*0.626*	0.619	0.498

Figure 4: Evaluation of classifiers by ROC curves and ROC-AUC. Models were trained on different inputs (titles or full texts), which were represented by different language models (columns).

5 Experiments

Classifiers were evaluated in the test set by area under the receiver operating characteristic curve (ROC-AUC), which assesses how well they balance sensitivity and specificity across a range of decision thresholds. This metric is appropriate because the frequency of coordinate occurrences ranges from the majority of articles to just dozens depending on the brain structure. Performance was summarized for each one-versus-all classification model by macro-averaging the ROC-AUC across brain structure classes. Superior performance was observed for classifiers taking term occurrences and GloVe embeddings as features (Figure 4). Across models using these features, training on the full texts offered uniformly improved performance. By contrast, the LSTM model performed at chance level on the full texts and did slightly better when trained on article titles.

6 Results and Discussion

By contrast to prior work that better predicted brain activity using LSTM encodings [4], our results support averaged GloVe embeddings as the language model of choice for this task. A key difference from the work by Jain *et al.* is that our text inputs have a long-range relationship with brain data. While their study related continuous natural speech inputs to evolving brain signals, our task was to relate a chunk of text thousands of words long to a single set of brain activation coordinates. The hypothesis that LSTM performs better on shorter strings of text is supported by its relatively better performance on article titles. This issue with LSTM may be exacerbated by vanishing/exploding gradient issues, though in this case, models were trained with gradient clipping.

Both the occurrence and GloVe models learned features that would be intuitive to a neuroscientist. In Figure 5, terms are sized by the probability with which they predicted activation when forward propagated through the classifiers for two selected brain regions. The left posterior middle temporal gyrus in particular has been associated with semantic processing of language, and related terms are strongly predictive of it across the occurrence and GloVe models.

Figure 5: Top 15 terms that most strongly predicted activation in a select brain structure.

7 Conclusions and Future Work

Our results point to GloVe as a low-dimensional summary of an article's semantic content that is relatively robust against overfitting. The model that used averaging performed only slightly better than that using a convolutional layer, which has potential to perform better with more exhaustive tuning of the convolution filter size, stride, and number of layers. Indeed, both the LSTM and convolutional GloVe model have greater potential for expressivity, and may be found to outperform the simpler models when the correct balance with regularization and dropout is achieved.

Another important direction for research will be to leverage this classification approach in experiments aimed at evaluating mental functions by their relation to brain activity. For example, ontologies of mental functions may be represented by text-based features and compared against one another by predictive performance. The development of methods here is intended as a first step for future approaches in computational neuroscientific ontology.

Contributions

Ellie collected and preprocessed the data, ran models in PyTorch, analyzed results in Python, and composed the manuscript and poster.

Code

The codebase for this project is available at `github.com/ehbeam/text2brain`.

Acknowledgments

Thank you to Suvadip Paul for his guidance. Acknowledgments to my advisor Amit Etkin, and to Russ Poldrack and Chris Potts for their mentorship on data processing and machine learning.

References

[1] Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., & Wager, T.D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665-670 (2011).

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

[3] Pennington, J., Socher, R., & Manning, C. GloVe: global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543 (2014).

[4] Jain, S. & Huth, A.G. Incorporating context into language encoding models for fMRI. *Conference on Neural Information Processing Systems* **32**, 1-10 (2018).

[5] Fox, P.T. & Lancaster, J.L. Mapping context and content: the BrainMap model. *Nat. Rev. Neurosci.* **3**, 319-321 (2002).

[6] Poldrack, R.A. *et al.* The Cognitive Atlas: toward a knowledge foundation for cognitive neuroscience. *Front. Neuroinform.* **5**, 1-11 (2011).

[7] Turner, J.A., & Laird, A.R. The Cognitive Paradigm Ontology: design and application. *Neuroinformatics* **10**, 57-66 (2012).

[8] Hastings, J., Ceusters, W., Jensen, M., Mulligan, K., & Smith, B. Representing mental functioning: ontologies for mental health and disease. *Third International Conference on Biomedical Ontology*, 1-5 (2012).

[9] Bug, W.J. *et al.*. The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, **6**, 175-194 (2008).

[10] Yarkoni, T. *Automated Coordinate Extractor (ACE)*. (GitHub, 2015).

[11] Lancaster, J.L. *et al.* Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Human Brain Mapping* **28**, 1194-1205 (2007).

[12] Desikan, R.S. *et al*. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968-980 (2006).

[13] Diedrichsen, J., Balster, J.H., Cussans, E., & Ramnani, N. A probabilistic MR atlas of the human cerebellum. *NeuroImage* **46**, 39-46 (2009).