fMRI Semantic Category Decoding using Linguistic Encoding of Word2Vec

Subba Reddy Oota
IIIT Hyderabad
oota.subba@students.iiit.ac.in

Naresh Manwani
IIIT Hyderabad
naresh.manwani@iiit.ac.in

Raju S. Bapi
IIIT Hyderabad
University of Hyderabad
India
raju.bapi@iiit.ac.in

Abstract

The dispute of how the human brain represents conceptual knowledge has been argued in many scientific fields. Brain imaging studies have shown that the spatial patterns of brain neural activation are correlated with thinking about different semantic categories of words (for example, tools, animals, and buildings) or when viewing the related pictures. In this paper, we present a computational model that learns to predict the neural activation captured in functional magnetic resonance imaging (fMRI) of test words. This model is trained with a combination of features extracted from the popular linguistic encoding of Word2Vec and the empirical fMRI data associated with viewing several dozen concrete nouns. We compared this model with several other models in which word features generated from FastText, Randomly, trillion-word text corpus. The experimental results show that the predicted fMRI images using Word2Vec meet the state of the art. The proposed scheme that uses popular linguistic encoding offers a simple and easy approach for semantic decoding from fMRI experiments.

1 Introduction

There are many scientific communities which studied how a human brain represents and organizes conceptual knowledge in the literature Caramazza & Mahon (2003); Mahon & Caramazza (2011); Tranel et al. (1997); Tong & Pratte (2012). In recent studies, the topic of exploring semantic representation in the human brain has attracted the attention of researchers from both neuroscience and computational linguistic fields. Using brain imaging studies Neuroscientists have shown that distinct spatial / temporal patterns of fMRI activity are associated with different stimuli such as face or scrambled face Clark et al. (1998), semantic categories of pictures, including tools, animals, and buildings, playing a movie etc Howell (2012); Haxby et al. (2001); Ishai et al. (1999); Kanwisher et al. (1997); Carlson et al. (2003); Cox & Savoy (2003); Haxby et al. (2000); Polyn et al. (2005). These experimental results postulate how the brain encodes meaning of words and knowledge of objects, including theories that meanings are encoded in the sensory-motor cortical areas Caramazza & Shelton (1998); Crutch & Warrington (2003); Samson & Pillon (2004). Such findings would also facilitate making predictions about breakdowns in the function and their spatial location in different neurological disorders. Theoretical and empirical studies have been conducted to explore animate and inanimate categorization and the brain representation of these semantic differences Cree & McRae (2003); Mahon & Caramazza (2005). Linguists have identified different

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

semantic meanings corresponding to individual verbs as well as the type of nouns that can fill those semantic meanings [e.g., WordNet Miller et al. (1990), VerbNet Schuler (2005), BableNet Navigli & Ponzetto (2010)]. Mitchell et al. (2008) pioneered studies that demonstrated common semantic representation for various nouns in terms of shared brain activation patterms across subjects.

At the heart of many of the fMRI decoding studies is the establishment of a mapping of the linguistic representation of nouns or verbs and the brain activation patterns elicited when subjects view these lexical items. One of the recent, popular approaches for linguistic representation of lexical items in computational linguistics is through a dense, low-dimensional and continuous vector called, word-embedding Hinton et al. (1986); Turney & Pantel (2010). Common word embeddings are generated from large text corpus like Wikipedia and statistics concerning the co-occurrence of words is estimated Mikolov et al. (2013); Bojanowski et al. (2016). Mitchell's team designed a computational model to predict the brain responses using hand-crafted word vectors as input to map the correlation between word embeddings and brain activity involved in viewing words Mitchell et al. (2008). Although these hand-crafted word-vectors cover different regions of the brain, the ambiguity between the word meanings are not covered. For example, the lexical item "Bank" has multiple semantic senses, such as the "bank of a river" or a "financial institution" based on the context. The recent popular approach FastText Bojanowski et al. (2016) is a fast and effective method to learn word representations and utilized for text classification. Since FastText embeddings are trained for understanding morphological variations and most of the syntactic analogies are morphology-based, FastText embeddings do significantly better on the syntactic analogies than on semantic tasks Bojanowski et al. (2016). To assess the relative strengths and weaknesses of these schemes, in this paper we made a comparison study to explore the correlation between brain activity involved in viewing words and different word embeddings including the popular approach of Word2Vec embedding Mikolov et al. (2013). To the best of our knowledge, this is the first time a comparative study is made among various exisiting, popular word embeddings for decoding brain activation. We propose a three-layer neural network architecture in which the input is a word embedding vector and the target output is the fMRI image depicting brain activation corresponding to the input word in line with the state of the art approaches Mitchell et al. (2008).

The structure of the paper is as follows. Section 2 describes the approach we are using to build the model, while Section 3 describes results of comparison of performance of various models along with the statistical significance of the results. In Section 4 we describe the conclusion and future work we are currently engaged in.

2 Approach

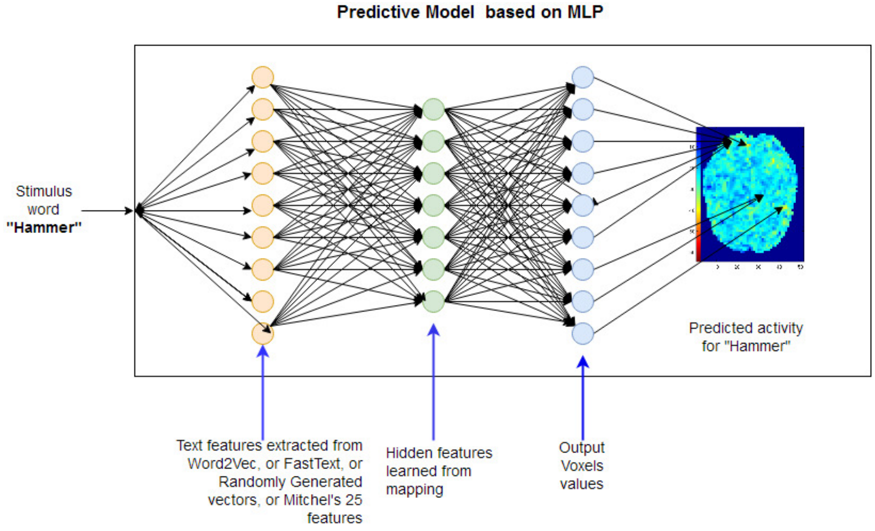
In this paper, we use the most common neural network architecture called the multilayer perceptron (MLP) Haykin (2001, chap. 4) to build a trainable computational model that predicts the neural activation for any given stimulus word **(w)** using a three-step process, shown in Fig. 1. Given a random stimulus word **(w)**, the first step is to provide the meaning of the word **(w)** as a vector of input features generated from one of the four different methods, such as Word2Vec, FastText, Randomly generated, or occurrences of stimulus word **(w)** within a large text corpus (25) used in Mitchell et al. (2008). Table 1 shows input feature vector for the word 'celery' using these four methods. The second step involves hidden layer representation and is accomplished via N hidden neurons in the hidden layer. Hidden neurons are fully connected to the input layer and the connection weights are learned through an adaptation process. The third step predicts the neural fMRI activation at every voxel location in the brain as a weighted sum of neural activations contributed by each of the hidden layer neurons. We present a comparative study using the four word encoding methods and compare relative accuracies and their statistical significance.

More precisely, the predicted activation z_v at voxel v in the brain for word w is given by

$$z_v = \sum_{j=1}^{N} c_{vj} f(net_j) + c_{j0} \tag{1}$$

$$f(net_j) = tanh(\sum_{i=1}^{M} c_{ij} x_i + c_{i0}) \tag{2}$$

Figure 1: **Multilayer Perceptron Architecture for decoding fMRI brain activation**


Predictive Model based on MLP
Stimulus word "Hammer"
Predicted activity for "Hammer"
Text features extracted from Word2Vec, or FastText, or Randomly Generated vectors, or Mitchel's 25 features
Hidden features learned from mapping
Output Voxels values

where, $f(net_j)$ is the value of the j^{th} hidden neuron for word w, N is the number of hidden neurons present in the model, and c_{vj} is a learned coefficient that specifies the degree to which the j^{th} intermediate semantic feature activates a voxel in the output layer.

3 Results

We experimented with this computational model using CMU fMRI data[1] of nine healthy subjects. These nine healthy subjects viewed 60 different word-picture pairs six times each. The 60 arbitrary stimuli included five items from each of the 12 semantic categories – animals, body parts, building parts, buildings, furniture, clothing, insects, kitchen items, tools, vegetables, vehicles, and other man-made items. For each stimulus, we computed a mean fMRI image over its six repetitions and the mean of all 60 of these stimuli was then subtracted to get the final representation image.

To begin our modeling exercise, we initially chose an input layer with input semantic features of each stimulus generated from either of Word2Vec, FastText, Randomly generated vectors, or Mitchell's 25 features. The reason behind the use of Random vectors is to see how a neural network represents brain activity even when provided with Random vectors at input layer, this experiment acting as a baseline control study. One of the important observations from Table 1 is that Mitchells' sematic feature vector would be expected to associate with stimulus-specific regions of the brain as for example, the feature 'see' would be related to the visual cortex, the feature 'say' would be related to the auditory cortex, the feature 'eat' to the sensory cortex, etc. Whereas, the word embedding method such as Word2Vec encodes the meaning in terms of co-occurrence frequencies of other words in the corpus and thus may not relate to various modules of the brain the way Mitechell's hand-crafted features are designed. However, the similarity score between the Word2Vec embedding vector for word "celery" and that of the Mitchell's 25 feature vector is approximately similar. In this way, even Word2Vec seems to capture the meaning in a way similar to Mitchell's scheme and perhaps might learn to elicit appropriate brain activation. Table 2 shows parameter settings for MLP model.

We trained separate computational models for each of the 9 participants using all the four input encoding methods. Each trained model was evaluated by means of a "leave-one-out" cross-validation approach in which the model was repeatedly trained with only 59 of the 60 available word stimuli and associated fMRI images. Each trained model was then tested by requiring that it first predict the fMRI image for the one "held-out" word. Figures 2, 3, and4 show ground truth fMRI image and the predicted fMRI image using the four methods for the words "bell", "arm", and "bee". It

[1] Available at `http://www.cs.cmu.edu/~fmri/science2008/data.html`

Word	Word2Vec	Word2Vec Similarity with Mitchells' 25	Mitchells' 25	Random	FastText
Celery	broccoli 0.705	eat 0.35	eat 0.837	0.89947	cabbage 0.74
	bell_peppers 0.697	taste 0.24	taste 0.346	-0.0600	carrots 0.74
	parsley 0.692	fill 0.0512	fill 0.315	-0.0600	onions 0.73
	cilantro 0.689	see 0.063	see 0.243	0.3208	spinach 0.73
	cabbage 0.688	clean 0.0549	clean 0.115	-0.2231	garlic 0.72
	cauliflower 0.675	open 0.042	open 0.060	-1.2333	tomato 0.70
	tomato 0.675	smell 0.189	smell 0.059	-0.1820	potatoes 0.70
	lettuce 0.673	touch 0.0618	touch 0.029	-0.1469	parsnips 0.69
	cherry_tomatoes 0.669	say 0.094	say 0.016	-0.3220	sweetroot 0.69
	Brussels_sprouts 0.669	hear 0.021	hear 0.000	-0.3220	lemongrass 0.69

Table 1: Top 10 features for the word Celery generated from the four methods

Parameters	Values
Hidden layer size	100
Optimizer	Adam
Activation	Tanh
Momentum, Learning rate	0.9, 0.001

Table 2: MLP parameter setting

can be observed from the Figures 2, 3, and 4 that Word2Vec and Mitchell's predicted fMRI images look visually similar to the actual fMRI image obtained during the empirical experiment, whereas Random and FastText results differ significantly.

We use the rescaled mean squared error (R^2) as a metric to measure the error between predicted and target fMRI brain images. Kruskal-Wallis rank test Kachigan (1986) was used for comparing mean ranks across the four methods. The one-way ANOVA test confirmed that there was a statistically significant difference between Word2Vec, FastText, and Randomly generated vectors. Table 3 shows a mean rank of 0.2963 for Word2Vec, 0.3160 for Mitchells', -0.0600 for Random, and -0.0089 for FastText. The post-hoc ScheffeâĂŹs test Scheffe (1953) results in Table 3 show that R^2 of Word2Vec differs significantly from that of the FastText vectors at p=0.001, Random vectors at p=0.001. No significant differences were observed between mean ranks of the Word2Vec and Mitchells' 25 features.

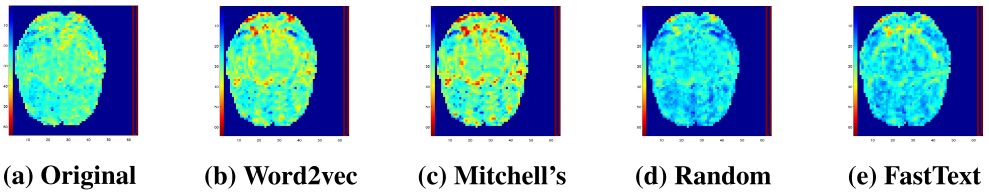
#Subject	Word2Vec	Mitchell's 25	Random	FastText	F statistic	p-value
Subj-1	0.2867	0.2765	-0.05600	-0.0078	13.2566	4.3283e-08**
Subj-2	0.2963	0.3169	-0.0600	-0.0089	14.9700	5.9377e-09**
Subj-3	0.2963	0.2924	-0.0600	-0.0089	13.3499	4.4471e-08**
Subj-4	0.4327	0.4273	0.3208	0.3435	9.1399	9.5623e-06**
Subj-5	0.1918	0.1800	-0.2231	-0.5236	30.1333	1.1102e-16**
Subj-6	-0.8066	-0.8213	-1.2333	-1.4631	1.7570	0.1561
Subj-7	0.2015	0.1896	-0.1820	-0.1564	13.5018	3.6773e-08**
Subj-8	0.2270	0.2200	-0.1469	-0.1710	29.8692	2.2204e-16**
Subj-9	0.1816	0.1751	-0.3220	-0.2670	15.0325	5.4972e-09**

**p<0.005

Table 3: UNIVARIATE TESTS FOR TESTING OF EQUALITY OF GROUP MEANS

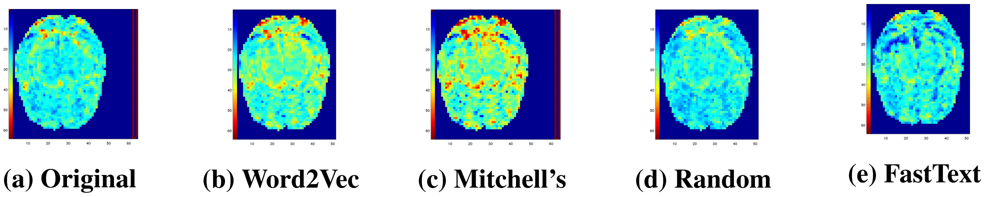
4 Conclusion

This study employs the existing popular word embeddings such as word2Vec, FastText to scrutinize the semantic representations in brain activity as measured by fMRI. One of the main observations from our study was that while Mitchells' hand-crafted features were designed to cover multimodal activity of the brain covering several brain regions, the corpus-based word embedding models are based on word co-occurrence based statistics and thus lack the multi-modal context embedded in Mitchell's feature vector. Such general word embedding encoding schemes tend to give a strong



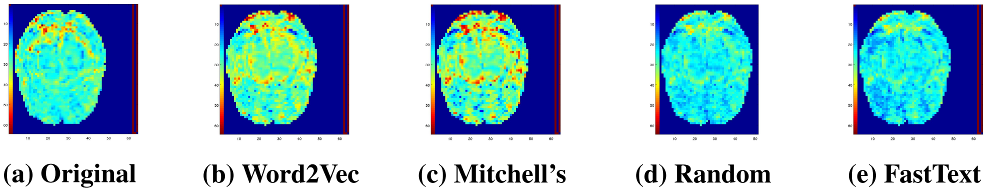
(a) Original (b) Word2vec (c) Mitchell’s (d) Random (e) FastText

Figure 2: Predicting fMRI images for given stimulus word "bell"



(a) Original (b) Word2Vec (c) Mitchell’s (d) Random (e) FastText

Figure 3: Predicting fMRI images for given stimulus word "arm"



(a) Original (b) Word2Vec (c) Mitchell’s (d) Random (e) FastText

Figure 4: Predicting fMRI images for given stimulus word "bee"

	Subjects (significance)								
Post-hoc	Subj-1	Subj-2	Subj-3	Subj-4	Subj-5	Subj-6	Subj-7	Subj-8	Subj-9
(1) vs (2)	0.89947	0.89947	0.89947	0.89947	0.89947	0.89947	0.89947	0.89947	0.89947
(1) vs (3)	0.001**	0.001**	0.001**	0.001**	0.001**	0.58985	0.001**	0.001**	0.001**
(1) vs (4)	0.237	0.001**	0.001**	0.005**	0.001**	0.22744	0.001**	0.001**	0.001**
(2) vs (3)	0.001**	0.001**	0.001**	0.0107*	0.001**	0.61378	0.001**	0.001**	0.001**
(2) vs (4)	0.001**	0.001**	0.001**	0.001**	0.001**	0.24580	0.001**	0.001**	0.001**

(1) Word2Vec features, (2)Mitchells’ 25 features , (3) Random features, (4) FastText features
**p<0.005, *p<0.05

Table 4: POST-HOC MULTIPLE COMPARISONS

within-category coverage for the input words and try to project this across different brain regions through associative mapping learned in the MLP. Thus the current study can be considered a feasibility study of using generic word embedding schemes for brain decoding. Experimental results reveal that the R^2 error between Mitchell’s approach and the proposed Word2Vec scheme is small and the statistical significance of the results also point out that both the approaches are similar in their final outcome. In future, we would like to generate word embeddings from corpora from different (multimodal) genres so that such feature vectors will also have an opportunity to learn mapping to multi-modal sensory and association regions of the brain. This might give us more insights into the mapping process of word-embedding representations to brain response and eventually improve the decoding accuracy of brain activation with such predictive solutions. source code publicly available at `https://github.com/subbareddy248/BrainDecoding` , so that the researcher and the developer communities can come forward, and collectively make the effort a grand success.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Alfonso Caramazza and Bradford Z Mahon. The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in cognitive sciences*, 7(8):354–361, 2003.

Alfonso Caramazza and Jennifer R Shelton. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of cognitive neuroscience*, 10(1):1–34, 1998.

Thomas A Carlson, Paul Schrater, and Sheng He. Patterns of activity in the categorical representations of objects. *Journal of cognitive neuroscience*, 15(5):704–717, 2003.

Vincent P Clark, Jose M Maisog, and James V Haxby. fmri study of face perception and memory using random stimulus sequences. *Journal of Neurophysiology*, 79(6):3257–3265, 1998.

David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri)âĂIJbrain readingâĂİ: detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003.

George S Cree and Ken McRae. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2):163, 2003.

Sebastian J Crutch and Elizabeth K Warrington. Spatial coding of semantic information: knowledge of country and city names depends on their geographical proximity. *Brain*, 126(8):1821–1829, 2003.

James V Haxby, Elizabeth A Hoffman, and M Ida Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233, 2000.

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Simon S Haykin. *Neural networks: a comprehensive foundation*. Tsinghua University Press, 2001.

Geoffrey E Hinton, James L Mcclelland, and David E Rumelhart. Distributed representations, parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations, 1986.

David C Howell. *Statistical methods for psychology*. Cengage Learning, 2012.

Alumit Ishai, Leslie G Ungerleider, Alex Martin, Jennifer L Schouten, and James V Haxby. Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 96(16):9379–9384, 1999.

Sam Kash Kachigan. *Statistical analysis: An interdisciplinary introduction to univariate & multivariate methods*. Radius Press, 1986.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

Bradford Z Mahon and Alfonso Caramazza. The orchestration of the sensory-motor systems: Clues from neuropsychology. *Cognitive neuropsychology*, 22(3-4):480–494, 2005.

Bradford Z Mahon and Alfonso Caramazza. What drives the organization of object knowledge in the brain? *Trends in cognitive sciences*, 15(3):97–103, 2011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4): 235–244, 1990.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 216–225. Association for Computational Linguistics, 2010.

Sean M Polyn, Vaidehi S Natu, Jonathan D Cohen, and Kenneth A Norman. Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756):1963–1966, 2005.

Dana Samson and Agnesa Pillon. Orthographic neighborhood and concreteness effects in the lexical decision task. *Brain and language*, 91(2):252–264, 2004.

Henry Scheffe. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2): 87–110, 1953.

Karin Kipper Schuler. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005.

Frank Tong and Michael S Pratte. Decoding patterns of human brain activity. *Annual review of psychology*, 63:483–509, 2012.

Daniel Tranel, Hanna Damasio, and Antonio R Damasio. A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia*, 35(10):1319–1327, 1997.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.