From brain space to distributional space: the perilous journeys of fMRI decoding

Gosse Minnema and Aurélie Herbelot
Center for Mind/Brain Sciences
University of Trento
gosseminnema@gmail.com
aurelie.herbelot@unitn.it

Abstract

Recent work in cognitive neuroscience has introduced models for predicting distributional word meaning representations from brain imaging data. Such models have great potential, but the quality of their predictions has not yet been thoroughly evaluated from a computational linguistics point of view. Due to the limited size of available brain imaging datasets, standard quality metrics (e.g. similarity judgments and analogies) cannot be used. Instead, we investigate the use of several alternative measures for evaluating the predicted distributional space against a corpus-derived distributional space. We show that a state-of-the-art decoder, while performing impressively on metrics that are commonly used in cognitive neuroscience, performs unexpectedly poorly on our metrics. To address this, we propose strategies for improving the model's performance. Despite returning promising results, our experiments also demonstrate that much work remains to be done before distributional representations can reliably be predicted from brain data.

1 Introduction

Over the last decade, there has been a growing body of research on the relationship between neural and distributional representations of semantics (e.g., Mitchell et al., 2008; Anderson et al., 2013; Xu et al., 2016). This type of research is relevant for cognitive neuroscientists interested in how semantic information is represented in the brain, as well as to computational linguists interested in the cognitive plausibility of distributional models (Murphy et al., 2012). So far, most studies investigated the correlation between neural and distributional representations either by predicting brain activity patterns from distributional representations (Mitchell et al., 2008; Abnar et al., 2018), or by using more direct correlation analyses like Representational Similarity Analysis (RSA; introduced in Kriegeskorte et al. 2008) or similar techniques (Anderson et al., 2013; Xu et al., 2016). Recently, however, a new model has been proposed (Pereira et al., 2018) for decoding distributional representations *from* brain images.

This new approach is different from the earlier approaches in a number of interesting ways. First of all, whereas predicting brain images from distributional vectors tells us something about how much neurally relevant information is present in distributional representations, doing the prediction in the opposite way could tell us something about how much of the textual co-occurrence information that distributional models are based on is present in the brain. Brain decoding is also interesting from an NLP point of view: the output of the model is a word embedding that could, at least in principle, be used in downstream tasks. Ultimately, a sufficiently accurate model for predicting distributional representations would amount to a sophisticated 'mind reading' device with numerous theoretical and practical applications.

Interestingly, despite being an early model and being trained on a (for NLP standards) very small dataset, Pereira et al. (2018) already report impressively high accuracy scores for their decoder. However, despite these positive results, there are reasons to doubt whether it is really possible to decode distributional representations from brain images. Given the high-dimensional nature of both neural and distributional representations, it is reasonable to expect that the mapping function between the two spaces, if it indeed exists, is potentially very complicated, and, given the inherent noisiness of fMRI data, could be very hard to learn, especially from a small dataset.

Moreover, we believe that the evaluation metrics used in Pereira et al. (2018) are too limited. Both of these metrics, *pairwise accuracy* and *rank*

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 155–161
Florence, Italy, July 28 - August 2, 2019.

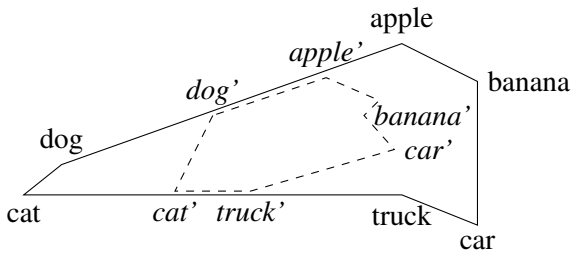

Figure 1: Hypothetical example where the predicted word embeddings (*cat', apple',* ...) are relatively close to their corresponding target word embeddings (cat, apple, ...), but are far from their correct position in absolute terms and have the wrong nearest neighbours.

accuracy, measure a predicted word embeddings's distance to its corresponding target word embedding, relative to its distance to other target word embeddings; for example, the prediction for *cat* is 'good' if it is closer to the target word embedding of *cat* than to the target word embedding of *truck* (see 3.1 for more details). Such metrics are useful for evaluating how well the original word labels can be reconstructed from the model's predictions, but do not say much about the overall quality of the predicted space. As shown in Figure 1, a bad mapping that fails to capture the similarity structure of the gold space could still get a high accuracy score. Scenarios like this are quite plausible given that cross-space mappings are known to be prone to over-fitting (and hence, poor generalization) and often suffer from 'hubness', a distortion of similarity structure caused by a lack of variability in the predicted space (Lazaridou et al., 2015).

In this paper, we fill a gap in the literature by proposing a thorough evaluation of Pereira et al. (2018), using previously untried evaluation metrics. Based on our findings, we identify possible weaknesses in the model and propose several strategies for overcoming these.

2 Related work

Our work is largely built on top of Pereira et al. (2018), which to date is the most extensive attempt at decoding meaning representations from brain imaging data. In this study (Experiment 1), fMRI images of 180 different content words were collected for 16 participants. The stimulus words were presented in three different ways: the written word plus an image representing the word, the word in a word cloud, and the word in a sentence. Thus, the dataset consists of $180 \times 3 = 540$ images per participant. Additionally, a combined representation was created for each word by averaging the images from the three stimulus presentation paradigms. Note that data for different participants cannot be directly combined due to differences in brain organization;[1] decoders are always trained for each participant individually.

The vocabulary was selected by clustering a pre-trained GloVe space (Pennington et al., 2014)[2] consisting of 30,000 words into regions, and then manually selecting a word from each region, yielding a set of 180 content words that include nouns (both concrete and abstract), verbs, and adjectives. Next, for every participant, a vector space was created whose dimensions are voxel activation values in that participant's brain scan.[3] This (approximately) 200,000-dimensional space can be optionally reduced to 5,000 dimensions using a complex feature selection process. Finally, for every participant, a ridge regression model was trained for mapping this brain space to the GloVe space. Crucially, this model predicts each of the 300 GloVe dimensions separately, the authors' hypothesis being that variation in each dimension of semantic space corresponds to specific brain activation patterns.

The literature relating distributional semantics to neural data started with Mitchell et al. (2008), who predicted fMRI brain activity patterns from distributional representations for 60 hand-picked nouns from 12 different semantic categories (e.g. 'animals', 'vegetables', etc.). Many later studies built on top of this; for example, Sudre et al. (2012) was a similar experiment using MEG, another neuroimaging technique. Other studies (e.g., Xu et al. 2016) reused Mitchell et al. (2008)'s original dataset but experimented with different word embedding models, including distributional models such as word2vec (Mikolov et al., 2013) or GloVe, perceptual models (Anderson et al., 2013; Abnar et al., 2018) and dependency-based models (Abnar et al., 2018). Similarly, Gauthier and Ivanova (2018) reused Pereira et al. (2018)'s data and regression model but tested it on alternative sentence embedding models.

[1] Techniques like *hyperalignment* do allow for this, but they require very large datasets (Van Uden et al., 2018).

[2] Version `42B.300d`, obtained from `https://nlp.stanford.edu/projects/glove/`.

[3] A voxel is a 3D pixel representing the blood oxygenation level of a small part of the brain.

3 Methods

Our work builds on top of Experiment 1 in Pereira et al. (2018) (described above) and uses the same datasets and experimental pipeline. In this section, we introduce our evaluation experiments (3.1) and our model improvement experiments (3.2).[4] Unless indicated otherwise, our models were trained on averaged fMRI images, which Pereira et al. showed to work better than using images from any of the individual stimulus presentation paradigms.

3.1 Evaluation experiments

Our evaluation experiments consist of two parts: a re-implementation of the pairwise and rank-based accuracy scores methods used in Pereira et al. (2018) and the introduction of additional evaluation metrics.

Pairwise accuracy is calculated by considering all possible pairs of words (u, v) in the vocabulary and computing the similarity between the predictions (p_u, p_v) for these words and their corresponding target word embeddings (g_u, g_v). Accuracy is then defined as the fraction of pairs where 'the highest correlation was between a decoded vector and the corresponding text semantic vector' (Pereira et al., 2018, p. 11). Unfortunately, the original code for computing the scores was not published, but we interpret this as meaning that a pair is considered to be 'correct' iff $\max(r(p_u, g_u), r(p_v, g_v)) > \max(r(p_u, p_v), r(p_v, p_u))$, where $r(x, y)$ is the Pearson correlation between two vectors. That is, for each pair of words, all four possible combinations of the two predictions and the two golds should be considered, and the highest of the four correlations should be either between p_u and g_u or between p_v and g_v.

Rank accuracy is calculated by calculating the correlation, for every word in the vocabulary, between the predicted word embedding for that word and all of the target word embeddings, and then ranking the target word embeddings accordingly. The accuracy score for that word is then defined as $1 - \frac{\text{rank}-1}{|V|-1}$, where *rank* is the rank of the correct target word embedding (Pereira et al., 2018, p. 11). This accuracy score is then averaged over all words in the vocabulary. Rank accuracy is very similar to pairwise accuracy but is slightly stricter. Under pairwise evaluation, it is sufficient if, for any word pair under consideration (say, *cat* and *dog*), only one of the predicted vectors is 'good': as long as the correlation between p_{cat} and g_{cat} is higher than the other correlations, the pair counts as 'correct', even if the prediction for *dog* is far off. Suppose that *dog* were the only badly predicted word in the dataset, then one could theoretically still get a pairwise accuracy score of 100%. By contrast, under rank evaluation a bad prediction for *dog* would not be 'forgiven' and the low rank of *dog* would affect the overall accuracy score, no matter how good the other predictions were.

In order to evaluate the quality of the predicted word embeddings more thoroughly, one would ideally use standard metrics such as semantic relatedness judgement tasks, analogy tasks, etc. (Baroni et al., 2014). However, this is not possible due to the limited vocabulary sizes of the available brain datasets. Instead, we test under four additional metrics that are based on well-established analysis tools in distributional semantics and elsewhere but have not yet been applied to our problem. The first two of these measure directly how close the predicted vectors are in semantic space relative to there expected location, whereas the last two measure how well the similarity structure of the semantic space is preserved.

Cosine (Cos) scores are a direct way of measuring how far each prediction is from 'where it should be', using cosine similarity as this is a standard metric in distributional semantics. Given a vocabulary V and predicted word embeddings (p_w) and target word embeddings (g_w) for every word $w \in V$, we define the cosine score for a given model as $\frac{\sum_{w \in V} \text{sim}(p_w, g_w)}{|V|}$ (i.e., the cosine similarity between each prediction and its corresponding target word embedding, averaged over the entire vocabulary).

$\mathbf{R^2}$ scores are a standard metric for evaluating regression models, and are useful for testing how well the value of each individual dimensions is predicted (recall that the ridge regression model predicts every dimension separately) and how much of their variation is explain by the model. We use the definition of R^2 scores from the `scikit-learn` Python package (Pedregosa et al., 2011), which defines it as the total squared distance between the predicted values and the true values relative to the total squared distance of each

[4] A software toolkit for reproducing all of our experiments can be found at https://gitlab.com/gosseminnema/ds-brain-decoding.

prediction to the mean true value:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2}$$

where y is an array of true values and $\hat{y}$ is an array of predicted values. Note that R^2 is defined over single dimensions; in order to obtain a score for the whole prediction matrix, we take the average R^2 score over all dimensions. Scores normally lie between 0 and 1 but can be negative if the model does worse than a constant model that always predicts the same value regardless of input data.

Nearest neighbour (NN) scores evaluate how well the similarity structure of the predicted semantic space matches that of the original GloVe space. For each word in V, we take its predicted and target word embeddings, and then compare the ten nearest neighbours of these vectors in their respective spaces. The final score is the mean Jaccard distance computed over all pairs of neighbour lists: $\frac{\sum_{w \in V} J(P_{10}(p_w), T_{10}(t_w))}{|V|}$, where $J(S,T) = \frac{|S \cap T|}{|S \cup T|}$ is the Jaccard distance between two sets (Lescovec et al., 2014) and $P_n(v)$ and $T_n(v)$ denote the set of n nearest neighbours (computed using cosine similarity) of a vector in the prediction space and in the original GloVe space, respectively.

Representational Similarity Analysis (RSA) is a common method in neuroscience for comparing the similarity structures of two (neural or stimulus) representations by computing the Pearson correlation between their respective similarity matrices (Kriegeskorte et al., 2008). We use it as an additional metric for evaluating how well the model captures the similarity structure of the GloVe space. This involves computing two similarity matrices of size $V \times V$, one for the predicted space and one for the target space, whose entries are defined as $P_{i,j} = r(p_i, p_j)$ and $T_{i,j} = r(t_i, t_j)$, respectively. Then, the representational similarity score can be defined as the Pearson correlation between the two upper halves of each similarity matrix: $r(\text{upper}(P), \text{upper}(T))$, where $\text{upper}(M) = [M_{2,1}, M_{3,1}, \ldots, M_{n,m-1}]$ is the concatenation of all entries $M_{i,j}$ such that $i > j$.

3.2 Model improvement experiments

The second part of our work tries to improve on the results of Pereira et al. (2018)'s model, using three different strategies: (1) alternative regression models, (2) data augmentation techniques, and (3) combining predictions from different participants.

Ridge is the original ridge regression model proposed in Pereira et al. (2018). Ridge regression is a variant on linear regression that tries to avoid large weights (by minimizing the squared sum of the parameters), which is similar to applying weight decay when training neural networks; this is useful for data (like fMRI data) with a high degree of correlation between many of the input variables (Hastie et al., 2009). However, an important limitation is that, when there are multiple output dimensions, the weights for each of these dimensions are trained independently. This seems inappropriate for predicting distributional representations because values for individual dimensions in such representations do not have much inherent meaning; instead, it is the interplay between dimensions that encodes semantic information, which we would like to capture this in our regression model.

Perceptron is a simple single-layer, linear perceptron model that is conceptually very similar to Ridge, but uses gradient descent for finding the weight matrix. A possible advantage of this approach is that the weights for all dimensions are learned at the same time, which means that the model should be able to capture relationships between dimensions. The choice for a linear model is also in line with earlier work on cross-space mapping functions (Lazaridou et al., 2015). Like Ridge, Perceptron takes a flattened representation of the 5000 'best' voxels as input (see section 2). Best results were found using a model using cosine similarity as the loss function, Adam for optimization (Kingma and Ba, 2014), with a learning rate and weight decay set to 0.001, trained for 10 epochs.

CNN is a convolutional model that takes as input a 3D representation of the full fMRI image. Our hypothesis is that brain images, like ordinary photographs, contain strong correlations between spatially close pixels (or 'voxels', as they are called in the MRI literature) and could thus benefit from a convolutional approach. We kept the CNN model as simple as possible and included only a single sequence of a convolutional layer, a max-pool layer, and a fully-connected layer (with a ReLU activation function). Best results were found with the same settings as for Perceptron, and a convolutional kernel size of 3 and a pooling ker-

Model	Pair			Rank			Cos			R2			NN			RSA		
	I_B	I_A	A	I_B	I_A	A	I_B	I_A	A	I_B	I_A	A	I_B	I_A	A	I_B	I_A	A
Random	0.54	0.50	0.49	0.54	0.50	0.51	-0.04	-0.05	-0.05	-3.16	-3.19	-2.48	0.04	0.03	0.03	0.01	-0.00	-0.01
Ridge	0.86	0.76	0.93	0.84	0.73	0.91	0.14	0.09	0.22	-0.30	-0.46	-0.06	0.07	0.05	0.11	0.14	0.08	0.25
Ridge+exp2	0.89*	**0.81***	**0.94***	0.86*	**0.79***	**0.92***	0.16*	**0.11***	0.23*	-0.12*	-0.25*	-0.06*	**0.09***	**0.06***	**0.12***	0.18*	**0.13***	0.25*
Ridge+para	**0.90**	0.78	**0.94**	**0.88**	0.75	**0.92**	**0.18**	0.10	**0.24**	**-0.16**	**-0.24**	**-0.05**	**0.09**	0.05	**0.12**	**0.20**	0.11	**0.26**
Ridge+aug	0.87	0.77	0.94	0.86	0.75	0.91	0.16	0.10	**0.24**	-0.18	-0.26	-0.05	0.07	0.05	0.11	0.16	0.09	0.25
Perceptron	0.81	0.70	0.87	0.78	0.68	0.83	0.09	0.05	0.11	-0.75	-41.89	-2.64	0.05	0.04	0.07	0.09	0.05	0.16
CNN	0.72	0.59	0.76	0.70	0.60	0.76	0.07	0.04	0.12	-0.40	-1.02	-0.13	0.05	0.03	0.05	0.08	0.03	0.13

Table 1: Decoding performance of all models. I_B: score of the best individual participant; I_A: average score for individual participants; A: score for the combined (averaged) predictions from all participants. '*' indicates that the model was tested on a subset of participants due to missing data.

nel size of 10.

We also propose several strategies for making better use of available data. **+exp2** adds completely new data points from Experiment 2 in Pereira et al. (2018)'s study: fMRI scans of 8 participants (who also participated in Experiment 1) reading 284 sentences, and distributional vectors for these sentences, obtained by summing the GloVe vectors for the content words in each sentence. By contrast, **+para** and **+aug** add extra data for every word in the existing vocabulary, in order to force the model to learn a mapping between regions in the brain space and regions in the target space, rather than between single points. In **+para**, the model is trained on four fMRI images per word: one from each stimulus presentation paradigm (i.e., the word plus a picture, the word plus a word cloud or the word in a sentence, and the average of these). By contrast, under the standard approach, the model is trained on only one brain image for each word (either the image from one of the three paradigms or the average image). Finally, **+aug** adds data on the distributional side: rather than mapping each brain image to just its 'own' GloVe vector (e.g. the image for *dog* to the GloVe vector of *dog*), we map it to its own vector plus the six nearest neighbours of that vector in the full GloVe space (e.g. not only *dog* but also *dogs, puppy, pet, cat, cats,* and *puppies*).

A final experiment does not aim at enhancing the models' training data, but rather changes how the model's predictions are processed. In the brain decoding literature, models are usually trained and evaluated for individual participants. However, to make maximal use of limited training data, one would like to combine brain images from different participants, but as noted, this is not feasible for our dataset. Instead, we propose a simple alternative method for obtaining group-level predictions: we average the predictions from all of the models for individual participants to produce a single prediction for each stimulus word. We hypothesize that this can help 'smooth out' flaws in individual participants' models. To compare individual-level and group-level predictions, we calculate three different scores for each model: the highest score for the predictions of any individual participant (I_B), the average score for the predictions of all individual participants (I_A), and the score for the averaged predictions (A).

4 Results

The results of all models are summarized in Table 1.[5] All models beat a simple baseline model that predicts vectors of random numbers (except on the R^2 metric, where Perceptron performs below baseline). Performance on the Pair and Rank scores is generally good, but performance on the other metrics is much worse: Cos is very low and R^2 scores are negative, meaning that the predicted word embeddings are very far in semantic space from where they should be. Moreover, the low NN and RSA scores indicate that the model captures the similarity structure of the GloVe space only to a very limited extent. On the model improvement side, the alternative models Perceptron and CNN fail to outperform Ridge, while the data augmentation experiments do achieve slightly higher performance. Finally, combining predictions seems to be quite effective: the scores for the averaged predictions are better than those for any individual participant, reaching Pair and Rank scores of more than 0.90 and Cos, NN, and RSA scores of up to two times the averaged score for individual participants.

5 Discussion and conclusion

Our results show that none of our tested models learns a good cross-space mapping: the predicted

[5]MLP and Ridge were run with and without feature selection; table lists best results (MLP: with, Ridge: without).

semantic vectors are far from their expected location and fail to capture the target space's similarity structure. Meanwhile, excellent performance is achieved on pairwise and rank-based classification tasks, which implies that the predictions contain sufficient information for reconstructing stimulus word labels. These contradictory results suggest a situation not unlike the one sketched in Fig. 1. This means that from a linguistic point of view, early claims about the success of brain decoding techniques should be taken cautiously.

Two obvious questions are how such a situation can arise and how it can be prevented. First of all, it seems likely that there is simply not enough training data to learn a precise mapping; the results of our experiments with adding 'extra' data are in line with this hypothesis. Moreover, the fact that all vocabulary words are relatively far from each other could make the mapping problem harder. For example, the 'correct' nearest neighbours of *dog* are *pig*, *toy*, and *bear*; the model might predict *fish*, *play* and *bird*, which are 'wrong' but intuitively do not seem much worse. We speculate that using a dataset with a more diverse similarity structure (i.e. where each word is very close to some words and further from others) could help the model learn a better mapping. Yet another issue is contextuality: standard GloVe embeddings are context-independent (i.e. a given word always has the same representation regardless of its word sense and syntactic position in the sentence), whereas the brain images are not because they were obtained using contextualized stimuli (e.g. a word in a sentence). Hence, an interesting question is whether trying to predict contextualized word embeddings, obtained using more traditional distributional approaches (e.g. Erk and Padó, 2010; Thater et al., 2011) or deep neural language models (e.g. Devlin et al., 2018), would be an easier task. Finally, the success of our experiment with combining participants suggests that using group-level data can help overcome the challenges inherent in predicting corpus-based (GloVe) representations from individual-level (brain) representations.

Acknowledgments

The first author of this paper (GM) was enrolled in the European Master Program in Language and Communication Technologies (LCT) while writing the paper, and was supported by the European Union Erasmus Mundus program.

References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 57–66.

Andrew J. Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1960–1970.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Katrin Erk and Sebastian Padó. 2010. Examplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97.

Jon Gauthier and Anna Ivanova. 2018. Does the brain represent words? an evaluation of brain decoding studies of language understanding. *CoRR*, abs/1806.00591. ArXiv preprint, http://arxiv.org/abs/1806.00591.

Trevor Hastie, Robert Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, second edition, corrected 7th printing edition. Springer Series in Statistics. Springer.

Diederik P. Kingma and Jimmy Lei Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980. ArXiv preprint, http://arxiv.org/abs/1412.6980.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Angeliki Lazaridou, Georgina Dinu, and Marco Baroni. 2015. Hubness and polution: Delving into

cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 270–280.

Jure Lescovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of Massive Datasets*, 2nd edition. Cambridge University Press. Online version, http://www.mmds.org/#ver21.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781. http://arxiv.org/abs/1301.3781.

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

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 114–123. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(963).

Gustavo Sudre, Dean Pomerleau, Palatucci Mark, Leila Wehbe, Alona Fyshe, Riita Salmelin, and Tom Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62:451–463.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkel. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1134–1143.

Cara E. Van Uden, Samuel A. Nastase, Andrew C. Connolly, Ma Feilong, Isabella Hansen, M. Ida Gobbini, and James V. Haxby. 2018. Modeling semantic encoding in a common neural representational space. *Frontiers in Neuroscience*, 12.

Haoyan Xu, Brian Murphy, and Alona Fyshe. 2016. Brainbench: A brain-image test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2017–2021.