BrainBench: A Brain-Image Test Suite for Distributional Semantic Models

Haoyan Xu
University of Victoria
Victoria, BC, Canada
exu@uvic.ca

Brian Murphy
Queen's University Belfast
Belfast, Northern Ireland, UK
brian.murphy@qub.ac.uk

Alona Fyshe
University of Victoria
Victoria, BC, Canada
afyshe@uvic.ca*

Abstract

The brain is the locus of our language ability, and so brain images can be used to ground linguistic theories. Here we introduce BrainBench, a lightweight system for testing distributional models of word semantics. We compare the performance of several models, and show that the performance on brain-image tasks differs from the performance on behavioral tasks. We release our benchmark test as part of a web service.

1 Introduction

There is active debate over how we should test semantic models. In fact, in 2016 there was an entire workshop dedicated to the testing of semantic representations (RepEval, 2016). Several before us have argued for the usage of brain data to test semantic models (Anderson et al., 2013; Murphy et al., 2012; Anderson et al., 2015), as a brain image represents a snapshot of one person's own semantic representation. Still, testing semantic models against brain imaging data is rarely done by those not intimately involved in psycholinguistics or neurolinguistics. This may be due to a lack of familiarity with neuroimaging methods and publicly available datasets.

We present the first iteration of BrainBench, a new system that makes it easy to test semantic models using brain imaging data (Available at `http://www.langlearnlab.cs.uvic.ca/brainbench/`). Our system has methodology that is similar to popular tests based on behavioral data (see Section 2.2), and has the additional benefit of being fast enough to offer as a web service.

2 The Tasks

Here we outline the set of tasks we used to evaluate several popular Distributional Semantic (DS) models.

2.1 Brain Image Data

For BrainBench we use two brain image datasets collected while participants viewed 60 concrete nouns with line drawings (Mitchell et al., 2008; Sudre et al., 2012). One dataset was collected using fMRI (Functional Magnetic Resonance Imaging) and one with MEG (Magnetoencephalography). Each dataset has 9 participants, but the participants sets are disjoint, thus there are 18 unique participants in all. Though the stimuli is shared across the two experiments, as we will see, MEG and fMRI are very different recording modalities and thus the data are not redundant.

fMRI measures the change in blood oxygen levels in the brain, which varies according to the amount of work being done by a particular brain area. An fMRI image is a 3D volume of the brain where each point in the volume (called a voxel) represents brain activity at a particular place in the brain. In the fMRI dataset used here, each voxel represents a 3mm x 3mm x 5mm area of the brain. Each of the 60 words was presented 6 times in random order, for a total of 360 brain images. The number of voxels depends on the size and shape of a person's brain, but there are around 20,000 voxels per participant in this dataset.

MEG measures the magnetic field caused by

*Corresponding Author

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2017–2021,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

many neurons firing in the same direction at the same time. This signal is very weak, and so must be measured in a magnetically shielded room. The MEG machine is essentially a large helmet with 306 sensors that measure aspects of the magnetic fields at different locations in the brain. A MEG brain image is the time signals recorded from each of these sensors. Here, the sampling rate is 200 Hz. For each word, the MEG recording is 800ms long resulting in 306×160 data points. Each of the words was presented 20 times (in random order) for a total of 1200 brain images. For simplicity we will use the term "brain image feature" to refer to both fMRI voxels and MEG sensor/time points.

A non-trivial portion of our participants' brain activity may be driven by the low-level visual properties of the word/line-drawing stimulus, rather than by semantics. As there is a possibility of confounding visual properties with semantic properties, we have attempted to remove the activity attributable to visual properties from the brain images. In total we have 11 visual features which include things like the length of the word, the number of white pixels, and features of the line drawing (Sudre et al., 2012). To remove the visual stimulus' contribution to the signal, we train a regression model that predicts the signal in each brain image feature as a function of the 11 visual features. We then subtract the predicted value from the observed value of the brain image feature. This process is known as "partialling out" an effect. Thus, the signal that remains in the brain image will not be correlated with the visual stimuli, and should only be related to the semantics of the word itself (or noise).

Brain images are quite noisy, so we used the methodology from Mitchell et al. (2008) to select the most stable brain image features for each of the 18 participants. The stability metric assigns a high score to features that show strong self-correlation over presentations of the same word. We noticed that tuning the number of features to keep made little or no difference in the absolute ordering of the different DS models. Thus, we use the optimal number of features averaged over all 6 DS models described in Section 3: the top 13% of MEG sensor/time points, and 3% of fMRI voxels. Finally, we average all brain images corresponding to repetitions of the same word.

2.2 Behavioral Tasks

We include, for comparison, four popular word vector evaluation benchmarks.

MEN This dataset contains 3,000 word pairs, such that each word appears frequently in two separate corpora. Human participants were presented with two word pairs and asked to choose the word pair that was more related, resulting in a ranking of relatedness amongst word pairs (Bruni and Baroni, 2013).

SimLex-999 A word pairing task meant to specifically target similarity rather than the more broad "relatedness" (Hill et al., 2015).

WS-353-[SIM|REL] A set of 353 word pairs with relatedness ratings (Finkelstein et al., 2002). This dataset was subsequently split into sets where the pairs denote similarity and relatedness, named WS-353-SIM and WS-353-REL, respectively (Agirre et al., 2009).

3 Distributional Models

We test six semantic models against both the fMRI and behavioral datasets. The six models are:

Skip-gram: A neural network trained to predict the words before and after the current word, given the current word. We selected a model with 300 dimensions trained on the Google news corpus (Mikolov et al., 2013).

Glove: A regression-based model that combines global context information (term-document cooccurrence) with local information (small windows of word-word cooccurrence) (Pennington et al., 2014). This 300-dimensional model was trained on the Wikipedia and Gigaword 5 corpora combined.

RNN: A recurrent neural network with 640-dimensional hidden vectors. These models are trained to predict the next word in a sequence and have the ability to encode (theoretically) infinitely distant contextual information (Mikolov et al., 2011). The model was trained on broadcast news transcriptions.

Global: A neural network model that incorporates global and local information, like that of the Glove model (Huang et al., 2012). This model is our smallest, with dimension 50, and was trained on Wikipedia.

Cross-lingual: A tool that projects distributional representations from multiple language into a shared representational space (Faruqui and Dyer, 2014). Here we use the German-English model (512 dimensions), trained on the WMT-2011 corpus.

Non-distributional: This model is based solely on hand-crafted linguistic resources (Faruqui and Dyer, 2015). Several resources like WordNet (Fellbaum, 1998) and FrameNet (Baker et al., 1998) are combined to make very sparse word vector representations. Due to their sparsity, these vectors are of very high dimension $(171, 839)$. This is a particularly interesting model because it is not built from a corpus (unlike every other model in this list).

Note that we are not aiming to compare the goodness of any of these distributional models, as they are trained on different corpora with different algorithms. Instead, we wish to compare the patterns of performance on behavioral benchmarks to that of a brain-image based task.

4 Methodology

Each of the behavioral tasks included here assigns a similarity score to word pairs. For each DS model we calculate the correlation between the vectors for every pair of words in the behavioral datasets. We then calculate the correlation between the DS vector correlations and the behavioral scores.

We follow a very similar methodology for the brain image datasets. Let us represent each DS model with a matrix $X \in \mathbb{R}^{w \times p}$ where w is the number of words for which we have brain images (here $w = 60$), and p is the number of dimensions in a particular DS model. From X we calculate the correlation between each pair of word vectors, resulting in a matrix $C_{DS} \in \mathbb{R}^{w \times w}$.

Let us represent each participant's brain images with a matrix $Y \in \mathbb{R}^{w \times v}$ where v is the number of selected brain image features. From this matrix we calculate the correlation between each pair of brain images, resulting in a matrix $C_{BI} \in \mathbb{R}^{w \times w}$ (BI for brain image). This final representation is similar to the behavioral tasks above, but now we have a similarity measure for *every* pair of words in our dataset.

Here is where the evaluation for brain imaging tasks differs from the behavioral tasks. Instead of measuring the correlation between C_{BI} and C_{DS}, as is done in Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), we use the testing methodology from Mitchell et al. (2008), which we will refer to as the 2 vs. 2 test. The 2 vs. 2 test was developed to help detect statistically significant predictions on brain imaging data, and, compared to RSA, can better differentiate the performance of a model from chance. We perform a 2 vs. 2 test for all pairs of C_{DS} and C_{BI} (that is, for every pair of DS model and fMRI/MEG participant).

For each 2 vs. 2 test we select the same two words (rows) w_1, w_2 from C_{DS} and C_{BI}. We omit the columns which correspond to the correlation to w_1 and w_2, as they contain a perfect signal for the 2 vs. 2 test. We now have four vectors, $C_{DS}(w_1)$, $C_{DS}(w_2)$, $C_{BI}(w_1)$ and $C_{BI}(w_2)$, all of length $w - 2$. We compute the correlation (corr) between vectors derived from C_{DS} and C_{BI} to see if:

$$\text{corr}(C_{DS}(w_1), C_{BI}(w_1)) + \text{corr}(C_{DS}(w_2), C_{BI}(w_2))$$

(the correlation of correctly matched rows: w_1 to w_1 and w_2 to w_2) is greater than:

$$\text{corr}(C_{DS}(w_1), C_{BI}(w_2)) + \text{corr}(C_{DS}(w_2), C_{BI}(w_1))$$

(the correlation of incorrectly matched rows). If the correctly matched rows are more similar than incorrectly matched rows, then the 2 vs. 2 test is considered correct. We perform the 2 vs. 2 test for all possible pairs of words, for 1770 tests in total. The 2 vs. 2 accuracy is the percentage of 2 vs. 2 tests correct. Chance is 50%.

Our process of computing 2 vs. 2 accuracy over rows of a correlation matrix is different than the original methodology for these datasets (Mitchell et al., 2008; Sudre et al., 2012). Previous work trained regression models that took brain images as input and predicted the dimensions of a DS model as output. Training these regression models for all 1770 pairs of words takes hours to complete, whereas the test we suggest here is much faster, and the correlation matrices C_{BI} can be computed ahead of time. This makes the tests fast enough to offer as a web service. We hope our web offering will remove barriers to the wider adoption of brain-based tests from within the computational linguistics community.

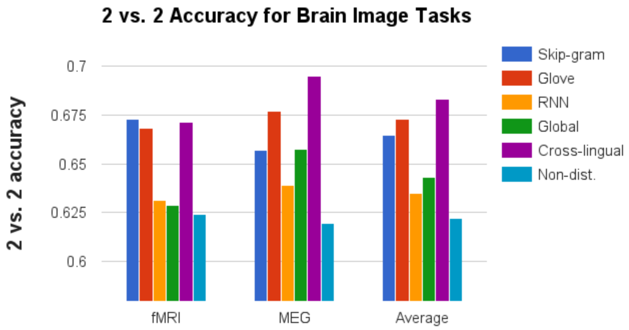

Figure 1: Performance of Distributional Semantic models on the brain-image datasets.

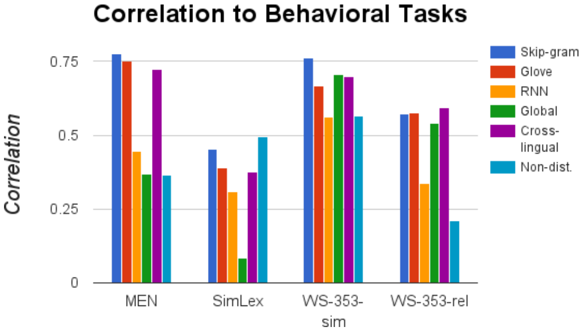

Figure 2: Performance of Distributional Semantic models on several benchmark behavioral tasks.

5 Results

Figure 1 shows the results for each of the DS models against the fMRI and MEG datasets. On average, the Skip-gram, Glove and Cross-lingual models perform quite well, whereas the multi-layer NNs (RNN, Global) perform less well. The one DS model to be built from hand-crafted resources (Non-distributional) performs poorly on both brain image tests.

As previously mentioned, we are not claiming to show that any one of the DS models is better than any other. Indeed, that would be comparing apples to oranges, as each DS model is trained with a different algorithm on a different corpus. Instead, notice that the *pattern* of performance for the fMRI task is remarkably similar to the pattern on the MEN behavioral task. This is interesting given that our dataset contains only 60 words and the MEN dataset contains > 700. On the MEG data, the Cross-lingual model performs best, and its performance pattern is unlike any of the behavioral tasks in Figure 2. The averaged BrainBench results are most similar to the results for WS-353-REL. However, averaging the results together may be misleading, as the fMRI and MEG result patterns are different.

6 Discussion

There are some caveats about the analyses herein. Firstly, the brain-based tests include only 60 concrete nouns, so they will necessarily favor distributional models with good noun representations, regardless of the representations of other parts of speech. We are currently working with various research groups to expand the number of brain-image datasets included in this benchmark to have a more diverse test base. The behavioral benchmarks were not reduced to include only the 60 words for which we have brain data, because this would have rendered the benchmarks essentially useless, as very rarely are a pair of the 60 words from the brain image data scored as a pair in the behavioral benchmarks.

7 Conclusion

We have presented our new system, BrainBench, which is a fast and lightweight alternative to previous methods for comparing DS models to brain images. Our proposed methodology is more similar to well-known behavioral tasks, as BrainBench also uses the similarity of words as a proxy for meaning. We hope that this contribution will bring brain imaging tests “to the masses” and encourage discussion around the testing of DS models against brain imaging data.

References

[Agirre et al.2009] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pas, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 19–27.

[Anderson et al.2013] 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 Conference on Empirical Methods on Natural Language Processing*.

[Anderson et al.2015] Andrew James Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. 2015. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, 120:309–322.

[Baker et al.1998] Collin F. Cf Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics* -, volume 1, page 86. Association for Computational Linguistics.

[Bruni and Baroni2013] Elia Bruni and Marco Baroni. 2013. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 48.

[Faruqui and Dyer2014] Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. *Proceedings of the European Association for Computational Linguistics*, pages 462–471.

[Faruqui and Dyer2015] Manaal Faruqui and Chris Dyer. 2015. Non-distributional Word Vector Representations. *Acl-2015*, pages 464–469.

[Fellbaum1998] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

[Finkelstein et al.2002] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

[Hill et al.2015] Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

[Huang et al.2012] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882.

[Kriegeskorte et al.2008] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(November):4, jan.

[Mikolov et al.2011] Tomáš Mikolov, Stefan Kombrink, Anoop Deoras, Lukáš Burget, and Jan Černocký. 2011. RNNLM — Recurrent Neural Network Language Modeling Toolkit. In *Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, pages 1–4.

[Mikolov et al.2013] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

[Mitchell et al.2008] 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 (New York, N.Y.)*, 320(5880):1191–5, may.

[Murphy et al.2012] Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 114–123, Montreal, Quebec, Canada.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe : Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar.

[RepEval2016] RepEval. 2016. RepEval workshop, ACL. `https://sites.google.com/site/repevalacl16/`.

[Sudre et al.2012] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking Neural Coding of Perceptual and Semantic Features of Concrete Nouns. *NeuroImage*, 62(1):463–451, may.