How the brain encodes meaning: Comparing word embedding and computer vision models to predict fMRI data during visual word recognition

Ning Mei (nmei@bcbl.eu)
Basque Center on Cognition, Brain, and Language
San Sebastian, Spain

Usman Sheikh (u.sheikh@bcbl.eu)
Basque Center on Cognition, Brain, and Language
San Sebastian, Spain

Roberto Santana (roberto.santana@ehu.eus)
Computer Science and Artificial Intelligence Department
University of Basque Country, Spain

David Soto (dsoto@bcbl.eu)
Basque Center on Cognition, Brain, and Language
San Sebastian, Spain

Abstract

The brain representational spaces of conceptual knowledge remain unclear. We addressed this question in a functional MRI study in which 27 participants were required to either read visual words or think about the concepts that words represented. To examine the properties of the semantic representations in the brain, we tested different encoding models based on word embeddings models -FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017), GloVe (Pennington, Socher, & Manning, 2014), word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)-, and, image vision models -VGG19 (Simonyan & Zisserman, 2014), MobileNetV2 (Howard et al., 2017), DenseNet121 (Huang, Liu, Van Der Maaten, & Weinberger, 2017)- fitted with the image referents of the words. These models were used to predict BOLD responses in putative substrates of the semantic network. We fitted and predicted the brain response using the feature representations extracted from the word embedding and computer vision models. Our results showed that computer vision models outperformed word embedding models in explaining brain responses during semantic processing tasks. Intriguingly, this pattern occurred independently of the task demand (reading vs thinking about the words). The results indicated that the abstract representations from the embedding layer of computer vision models provide a better semantic model of how the brain encodes word meaning. `https://tinyurl.com/y5davcs6`.

Keywords: fMRI, semantics, encoding model

Introduction

Semantic memory is defined as the cognitive function that holds and retrieves language related information (Binder, Desai, Graves, & Conant, 2009). fMRI-based classification studies have shown the semantic category of both pictures and words can be decoded from multivoxel patterns in different brain regions (Bauer & Just, n.d.) of the so-called semantic network. Encoding models further enable us to reach to a comprehensive level of how semantic information is represented during language processing tasks and define how the brain derives a cognitive map of meaning (Naselaris, Kay, Nishimoto, & Gallant, 2011; Felsen & Dan, 2006).

Word embedding algorithms (i.e. word embedding models (Mikolov et al., 2013)) have shown that semantic knowledge is organized in a meaningful way: words that share similar semantics tend to have closer distances in a high dimensional space that is defined by these algorithms, such as word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017). These models are usually 3-layer neural network models and the outputs (i.e. a vector of 300 elements) of the middle layer are used as the feature representation of a given word that exists in the model training corpus.

Studies using computer vision models (i.e. deep convolutional neural networks (LeCun, Bottou, Bengio, & Haffner, 1998)) have also shown the structural organization of meaning (Simonyan & Zisserman, 2014), which is image-based. For instance, a computer vision model trained on a set of images can learn the semantic categories defined by the target labels. The trained-model not only manages to generalize learned categories to unseen images but also can be used as a feature extractor to facilitate transfer learning in new classification tasks (Jia et al., 2014; Yosinski, Clune, Bengio, & Lipson, 2014). Features extracted by the second to last layer of the convolutional neural network, which is often the global pooling layer, are usually used as the feature representation of a given image (Jia et al., 2014).

In our fMRI experiment, we had 2 experimental contexts (i) participants were asked to read words that were visually presented (condition read) or (ii) to think about the properties and experiences associated with the words (condition think). We asked (i) whether the representations of the words in the human brain are related to those produced by the word embedding algorithms (ii) whether the strength of this relationship is dependent on the task context (i.e., reading vs thinking) and (iii) whether word embedding algorithms provide a better account of the brain representational spaces relative to computer vision models.

To investigate the associative mapping of the semantic representations of words and the corresponding brain activity patterns, we implemented different encoding models based on different word embeddings and image models.

An encoding model predicts the brain activity patterns using a set of features that are linearly or nonlinearly transformed from the stimuli (Diedrichsen & Kriegeskorte, 2017; Kriegeskorte & Douglas, 2018). In order to map the sensory stimuli to the brain activity patterns, the encoding model reconstructs the brain activity patterns by utilizing a given set of feature/representational space extracted from the stimuli (Naselaris & Kay, 2015).

We hypothesized image-like features were more likely to be mentally represented during the think task relative to the read task. Therefore, besides three word-embedding models, we selected three computer vision models to extract features from images corresponding to the words we used in the experiment. These computer vision models were VGG19 (Simonyan & Zisserman, 2014), MobileNetV2 (Howard et al., 2017), and DenseNet121 (Huang et al., 2017).

A set of 7 left-lateralized regions of interest (ROIs) in the well-known semantic network were pre-specified based on Binder et al. (2009), and they included inferior parietal lobe (IPL), lateral temporal lobe (LTL), ventromedial temporal lobe including fusiform gyrus and parahippocampal gyrus (FFG & PHG), dorsomedial prefrontal cortex (dmPFC), inferior frontal gyrus (IFG), ventromedial prefrontal cortex (vmPFC), and posterior cingulate gyrus (PCG) along with precuneus. We hypothesized that, in these regions, the word embedding features could explain the BOLD responses better than computer vision model features when participants were asked to read the words, while the computer vision model features explain the brain activity patterns better when participants were asked to think of related information of the words.

Method

Experimental procedure and Stimuli

There were two task conditions in the experiment (i) read and (ii) think, with 4 blocks of 36 trials in each condition. In the read condition, participants were asked to silently read the visual words that appeared on the computer screen. In the think condition, participants were asked to think about the words. There were 36 Spanish words and 18 of them were living things (i.e. tiger, horse) while the other 18 were nonliving (i.e. pencil).

fMRI acquisition and preprocessing

3-Tesla magnet and 64-channel head coil SIEMENS's Magentom Prisma-fit scanner was used to collect data. For each participant, one high-resolution T1-weighted structural image and 8 functional MRI sessions were acquired. In each fMRI session, a multiband gradient-echo echo-planar imaging sequence with an acceleration factor of 6, resolution of 2.4x2.4x2.4mm^2, TR of 850 ms, TE of 35 ms and bandwidth of 2582 Hz/Px was used to obtain 585 3D volumes of the whole brain (66 slices; FOV = 210mm).

Word Embedding Models

The word embedding models used in the experiment were pretrained (Bravo-Marquez & Kubelka, 2018) based on each of the corresponding proposed methods using the Spanish Billion Word Corpus[1]. For each word used in the experiment, the corresponding vector representation of 300 dimensions was extracted.

Computer Vision Models

The computer vision models used in the experiment were pretrained models provided by the Keras Python library(Chollet, 2015) based on each of the corresponding proposed methods using the ImageNet dataset[2]. For each word used in the experiment, we sampled 10 images collected from the internet. Images were cropped and the object appeared at the center on white background. The output vector of the global average layer of the computer vision model for a given image was its feature representation. The 10 vector representations corresponded to the same word were averaged. The feature extraction was done trial-wise accordingly for each participant, each ROI, and each experiment condition.

Encoding Model Pipeline

The encoding model pipeline was the same as in Miyawaki et al. (2008, also see nilearn (Pedregosa et al., 2011; Buitinck et al., 2013)). After standardizing the feature representations by subtracting the mean and divided by standard deviation, the feature representations were mapped to the BOLD signal of a given ROI by an L2 regularized regression (Ridge Regression) with a regularization term of 100.

To estimate the performance of the regression, we partitioned the data into 100 folds to perform cross-validation by stratified random shuffling (Little et al., 2017). In each fold, we randomly held out 20% of the data for testing, while the rest 80% were used to fit the ridge regression model, using the feature representations from word embedding models or computer vision models as features and the BOLD signals as targets, and then create the predictions for the held-out data. The percentage of variance explained in each voxel was computed for the predictions. An average estimate of the variance explained was calculated. Voxels that had positive variance explained values were kept for further analysis (Miyawaki et al., 2008).

For each participant, ROI and condition, we computed the average of the variance explained. To estimate the empirical chance level performance of the encoding models, a ran-

[1] http://crscardellino.github.io/SBWCE/

[2] http://www.image-net.org/

dom shuffling was added to the training phase during cross-validation before model fitting. The random shuffling was applied to the order of the samples for the features in the encoding models while the order of the samples for the targets remained the same.

Results

Figure 1 shows the average explained variance by each encoding model. The error bars represent bootstrapped 95% confidence interval across 27 participants.

The estimated empirical chance level performance of all the encoding models was zero: when the feature representation of a word and its corresponding trials were permuted by swapping the order of the feature representation matrix while keeping the order of samples of the BOLD signals, there was zero variance explained. This result suggests that variance explained estimated from a given set of feature representations was not due to chance level noise.

First, we conducted one-way ANOVAs separately for the word embedding models and the computer vision models in each ROI and each experiment condition. In the read condition, there was no difference among the word embedding models in terms of variance explained but there was a difference among the computer vision models with the mobilenet performing better than the others (Fig.1, upper panel). The same patterns held in the think condition for only FFG and PHG, PCG and Precun, and vmPFG (Fig.1, lower panel)

Then, we compared each pair of word embedding model and computer vision models. We subtracted the variance explained of the word embedding model from the computer vision model and then performed a permutation t-test against zero variance explained. Figure 2 shows that the computer vision models explained more variance across the participants than the word embedding models across the 'read' and 'think' conditions and in all ROIs (all ps < 0.05, Bonferroni corrected).

Finally, we computed the average of variance explained across the word embedding models or also across the computer vision models for each participant, each ROI, and each experimental condition. Then we tested for the pairwise differences between the two experimental conditions (read vs. think) for each ROI. Figure 3 shows the violin plots of these differences of each experimental condition. The ranges of the the violin figures represent the minimums and maximums of the distribution, while the inner lines represents the 25%, 50%, and 75% quartile of the distribution. As showed in the figure, computer vision models significantly explained more variance than the word embedding models in the think condition compared to the read condition for the FFG and PHG as well as PCG and Precun.

Figures

Discussion

Overall we found that computer vision models outperformed word embedding models in explaining brain responses during semantic processing tasks. This pattern occurred independently of the task demand (reading vs thinking about the words). However, computer vision models predicted more variance in visual areas such as the fusiform in the think condition, which is consistent with participants accessing to visual representations during mental simulation of the concept. Intriguingly, even during the read condition the computer vision model was better at explaining the brain responses. These data indicate that the abstract representations from the embedding layer of computer vision models provide a better "semantic" model of how the brain encodes word meanings.

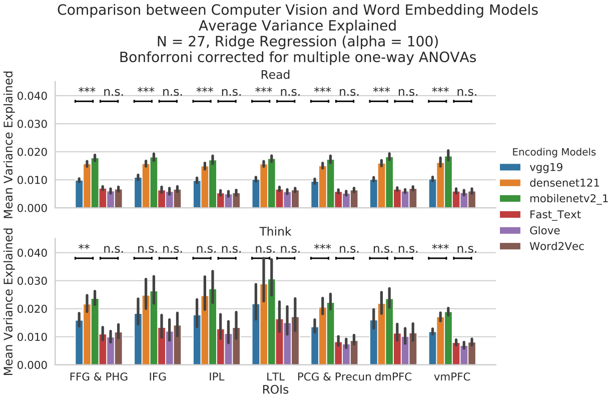

Figure 1: Average Variance Explained by each Encoding Model

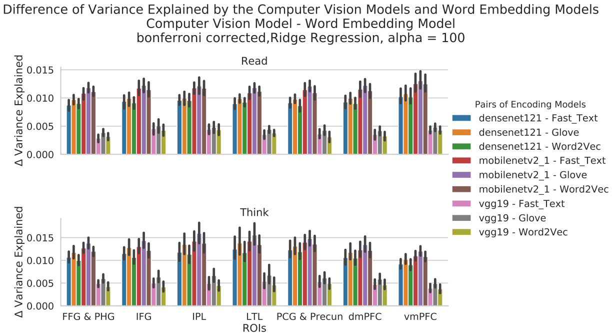

Figure 2: Differences between each Computer Vision Model and Word Embedding Model in Average Variance Explained

Acknowledgments

D.S. acknowledges support from the Spanish Ministry of Economy and Competitiveness, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R & D (SEV-2015-490) and project grants PSI2016-76443-P from MINECO and PI-2017-25 from the Basque Government.
R. Santana acknowledges support by the Basque Government (ELKARTEK programs), and Spanish Ministry of Economy and Competitiveness MINECO (project TIN2016-78365-R).

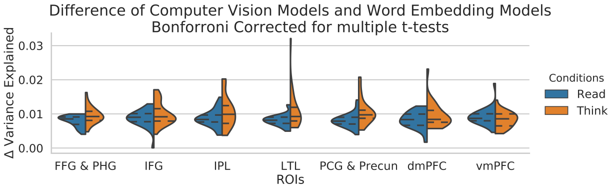

Figure 3: Differences between Read and Think Conditions

References

Bauer, A. J., & Just, M. A. (n.d.). Neural representations of concept knowledge. In *The Oxford Handbook of Neurolinguistics* (chap. 21).

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bravo-Marquez, F., & Kubelka, J. (2018). *spanish-word-embeddings.* `https://tinyurl.com/y47xzh6l`. GitHub.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., . . . Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML pkdd workshop: Languages for data mining and machine learning* (pp. 108–122).

Chollet, F. (2015). *keras.* `https://tinyurl.com/pna7m6p`. GitHub.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, *13*(4), e1005508.

Felsen, G., & Dan, Y. (2006, 01). A natural approach to study vision. *Nature Neuroscience*, *8*, 1643-6. doi: 10.1038/nn1608

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., . . . Darrell, T. (2014). Caffe. *Proceedings of the ACM International Conference on Multimedia - MM 14*. Retrieved from `https://tinyurl.com/y6qhfzsr` doi: 10.1145/2647868.2654889

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 1.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). Using and understanding cross-validation strategies. perspectives on saeb et al. *GigaScience*, *6*(5), gix020.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., . . . Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915–929.

Naselaris, T., Kay, K., Nishimoto, S., & Gallant, J. (2011, 05). Encoding and decoding in fmri. *NeuroImage*, *56*, 400-10. doi: 10.1016/j.neuroimage.2010.07.073

Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of mvpa using explicit models of representation. *Trends in cognitive sciences*, *19*(10), 551–554.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).