Comparing word embedding and computer vision models to predict fMRI data during visual word recognition

Ning Mei[1], **Usman Sheikh**[1], **Roberto Santana**[2], **David Soto**[1]

1. Basque Center on Cognition, Brain, and Language, San Sebastian, Spain
2. University of Basque Country, Spain

semsociaty, d.soto.b(@gmail.com)

Introduction

1. Semantic memory is the cognitive function that holds and retrieves language related information [2].
2. fMRI-based classification studies [1] have shown that the semantic category of both pictures and words can be decoded from multivoxel patterns in different brain regions of the so-called semantic network [2]. **Encoding models** further enable us to understand the type of information that is represented in brain activity during language processing tasks and define how the brain derives a cognitive map of meaning [11, 5].
3. However, how the brain representation of conceptual knowledge varies as a function of internal processing goals and strategies remains unclear.
4. Word embedding algorithms (i.e. Word2Vec-2013 [10]) provide a way to characterize the geometry of semantic spaces.
5. Computer vision models (i.e. deep convolutional neural network [9] also reveal the structural organization of meaning in the ventral visual pathway [14].
6. To examine the properties of the semantic representations in the brain during word recognition, we tested different encoding models based on **word embedding models** and **computer vision models**.

Experiment

1. 27 participants.
2. 18 living and 18 non-living **words**. Note: no image of objects was shown
3. The experiment was comprised of 8 fMRI runs. Each trial began with a fixation period of 250 ms followed by a blank screen of 500 ms (see Figure 1) and then by the target visual word which was displayed for **1 s**.
4. Depending on session instructions (shallow or deep processing), the participants were asked to either read and attend to the word (**shallow process condition**), or mentally simulate the properties associated with the word (e.g. its shape, its color etc., **deep process condition**).
5. Following a prior meta-analysis [2], 15 left-lateralized areas were pre-specified and included regions withing inferior parietal lobe, lateral temporal, ventromedial temporal lobe including fusiform gyrus and parahippocampal gyrus, dorsomedial prefrontal cortex, inferior frontal gyrus, ventromedial prefrontal cortex, posterior cingulate gyrus and anterior temporal lobe.
6. Word embedding models includes Word2Vec-2013 [10], GloVe [13], and Fast Text [3]
7. Computer vision models includes VGG-19 [14], DenseNet121 [6], and MobileNetV2 [7]

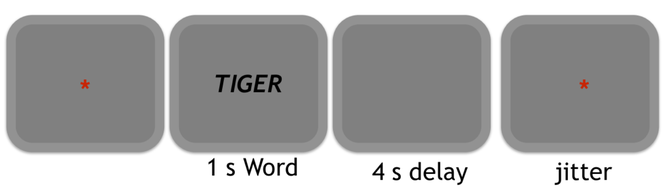

Figure 1: *Experiment Paradigm*

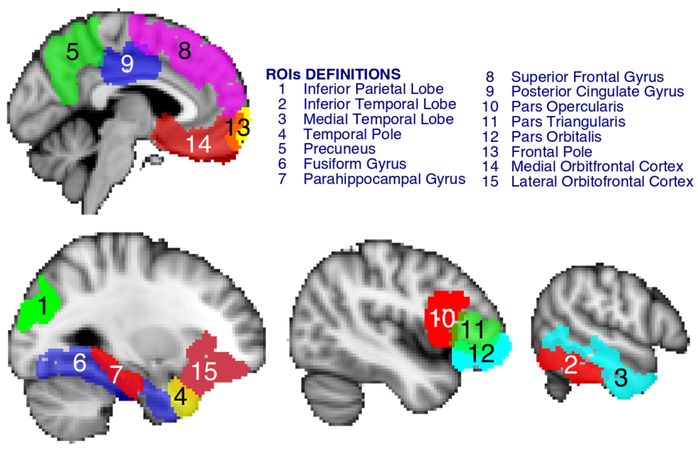

Figure 2: *Region of Interests*

Encoding Model

1. An encoding model predicts the brain activity patterns using a set of features that are (non)linearly transformed from the stimuli [4, 8]. In order to map the sensory stimuli to the brain activity, the encoding model reconstructs/predicts the brain activity patterns by utilizing a given set of feature/representational spaces extracted from the stimuli [12].
2. We hypothesized image-like features were more likely to be mentally represented during the 'deep information processing' condition relative to the 'shallow processing' condition. Therefore, besides three word-embedding models (Word2Vec-2013, GloVe, and Fast-Text), we selected three computer vision models (VGG-19, MobileNetV2, and DenseNet121) to extract features from images corresponding to the words we used in the experiment.

Results, ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, n.s.: not significant

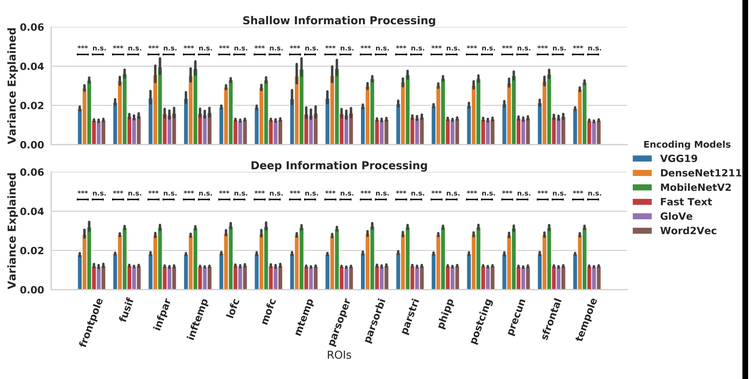

Figure 3: *Average Variance Explained*

We subtracted variance explained of each word embedding model from each computer vision model. And then we compared each pair of differences against zero. The multiple comparison was corrected with either shallow or deep process condition by FDR Benjamini-Hochberg procedure.

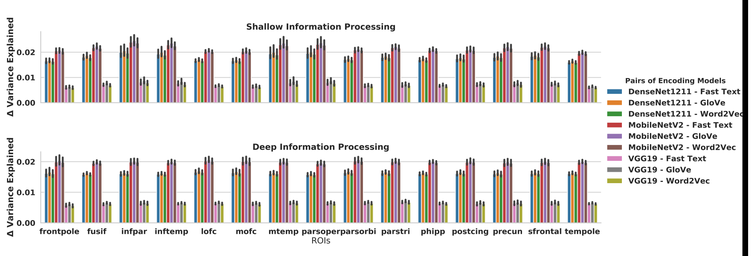

Figure 4: *Difference of Variance Explained by the Computer Vision Model and Word Embedded Model*

We then averaged the difference between each computer vision and word embedding model, for each subject and for both shallow or deep processing conditions, and compared these across subjects. Multiple comparison correction was performed across the ROIs using FDR Benjamini-Hochberg procedure.

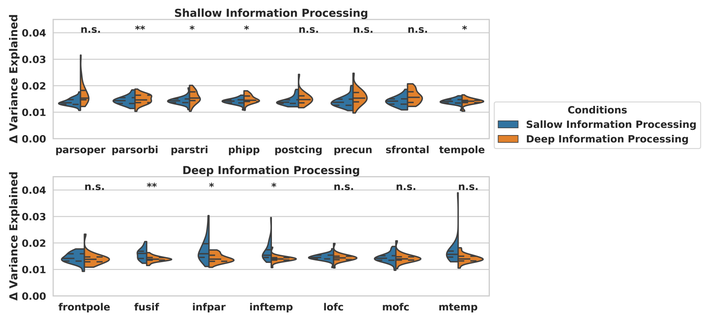

Figure 5: *Difference of Computer Vision Models and Word Embedding Models contrast between shallow and deep process*

Conclusion and Limitation

1. Computer vision models outperformed word embedding models in explaining brain responses during semantic processing tasks. This pattern occurred independently of the task demand (shallow vs deep process the words).
2. Computer vision models predicted more variance in visual areas such as the fusiform during deep information process condition, which is consistent with participants accessing to visual representations during mental simulation of the concept.
3. The abstract representations from the embedding layer of computer vision models provide a better "semantic" model of how the brain encodes word meanings [15].
4. The representations of the computer vision models are much larger than those of the word embedding models (1000^{+} v.s. 300). This is due to the fact that none of the models was fine-tuned before applied to the dataset.

References

[1] Andrew James Bauer and Marcel Adam Just. Neural representations of concept knowledge. In *The Oxford Handbook of Neurolinguistics*, chapter 21.

[2] Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796, 2009.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

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

[5] Gidon Felsen and Yang Dan. A natural approach to study vision. *Nature Neuroscience*, 8:1643–6, 01 2006.

[6] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[8] Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, page 1, 2018.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[10] 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*, pages 3111–3119, 2013.

[11] Thomas Naselaris, Kendrick Kay, Shinji Nishimoto, and Jack Gallant. Encoding and decoding in fmri. *NeuroImage*, 56:400–10, 05 2011.

[12] Thomas Naselaris and Kendrick N Kay. Resolving ambiguities of mvpa using explicit models of representation. *Trends in cognitive sciences*, 19(10):551–554, 2015.

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Julia Vogel and Bernt Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007.

Acknowledgements

D.S. and N.M. 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. S. acknowledges support by the Basque Government (ELKARTEK programs), and Spanish Ministry of Economy and Competitiveness MINECO (project TIN2016-78365-R).