

Interpretable Semantic Vectors from a Joint Model of Brain- and Text-Based Meaning Fyshe et al.

ABNS 2023

Vector Space Model (VSM)

Construction and use



Vector Space Model

- VSMs represent lexical meaning by assigning each word a point in high dimensional space. The high dimensional space can be any vectorial representation associated with each word.
- VSMs are *typically* created using a large text corpora. When this is the case, the VSM represents word semantics *as observed in text*
- In the VSM, the distance between any two words is taken to indicate their semantic similarity (matching, e.g., that observed and rated by speakers)
- Corpus-based VSMs have been criticized as being noisy or incomplete representations of meaning



VSM from brain activity?

- When a person is reading or writing, the semantic content of each word necessarily produces patterns of activity over neurons/voxels/sensors
- In principle then, brain activity could be used instead of corpus data to construct VSM.
- *If brain activation* data encodes semantics, including brain data in a model of semantics could result in a more effective model.
 - “The inclusion of brain data will only improve a text-based model if brain data contains semantic information not readily available in the corpus.”

Challenges of the project

Create

Create a database of human-annotated word semantics and create a brain-informed VSM that better predicts this database

Predict

Predict corpus representations of withheld words more accurately by infusing brain data into the learning model

Map

Map semantic concepts onto the brain by jointly learning neural representations

Data

Corpus and Brain data

Corpus Data

- The corpus data are compiled from a 16 billion word subset of ClueWeb09 and contain two types of corpus features: dependency and document features.
- Dependency statistics were derived by dependency parsing the corpus and compiling counts for all dependencies incident on the word.
- Count thresholding was applied to reduce noise, and positive pointwise mutual-information (PPMI) was applied to the counts. SVD was applied to the document.

Brain Data

- fMRI data and MEG data for 18 subjects (9 in each imaging modality)
- Each read 60 concrete nouns. The 60 words span 12 word categories

Method

NNSE and JNNSE

NNSE

$$\operatorname{argmin}_{A,D} \sum_{i=1}^w \|X_{i,:} - A_{i,:} \times D\|^2 + \lambda \|A\|_1 \quad (1)$$

$$\text{subject to: } D_{i,:} D_{i,:}^T \leq 1, \forall 1 \leq i \leq \ell \quad (2)$$

$$A_{i,j} \geq 0, \quad 1 \leq i \leq w, \quad 1 \leq j \leq \ell \quad (3)$$

Interpretability of A matrix

- What's the 'meaning' of a word?
 - Find top scoring dimensions for the word (i.e columns with highest values for the word's row).
 - Then, find words that score highest on those dimensions
- "For example, the word chair has the following top scoring dimensions:
 - 1. chairs, seating, couches;
 - 2. mattress, futon, mattresses;
 - 3. supervisor, coordinator, advisor. These dimensions cover two of the distinct meanings of the word chair (furniture and person of power)"

JNNSE: adding brain-activity information

Corpus stat per word

$$\begin{aligned} \operatorname{argmin}_{A, D^{(c)}, D^{(b)}} & \sum_{i=1}^w \|X_{i,:} - \boxed{A_{i,:}} \times D^{(c)}\|^2 + \\ & \sum_{i=1}^{w'} \|Y_{i,:} - \boxed{A_{i,:}} \times D^{(b)}\|^2 + \lambda \|A\|_1 \end{aligned} \quad (4)$$

Activity per word

$$\text{subject to: } D_{i,:}^{(c)} D_{i,:}^{(c)T} \leq 1, \forall 1 \leq i \leq \ell \quad (5)$$

$$D_{i,:}^{(b)} D_{i,:}^{(b)T} \leq 1, \forall 1 \leq i \leq \ell \quad (6)$$

$$A_{i,j} \geq 0, 1 \leq i \leq w, 1 \leq j \leq \ell \quad (7)$$

JNNSE


Corpus stat per word

$$\boxed{X} = \begin{bmatrix} w_0 & c_0 & \dots \\ \vdots & \vdots & \ddots \\ w_w & c_c \end{bmatrix}$$

As many words
as words in
corpus 'w'

$$A = \begin{bmatrix} w_0 & l_0 & \dots \\ \vdots & \vdots & \ddots \\ w_w & l_l \end{bmatrix}$$

$$\operatorname{argmin}_{A, D^{(c)}, D^{(b)}} \sum_{i=1}^w \|X_{i,:} - A_{i,:} \times D^{(c)}\|^2 + \sum_{i=1}^{w'} \|Y_{i,:} - A_{i,:} \times D^{(b)}\|^2 + \lambda \|A\|_1$$

$$D^{(c)} = \begin{bmatrix} c_0 & \dots \\ \vdots & \vdots \\ l_l \end{bmatrix}$$


JNNSE


Activity per word

$$\boxed{Y} = \begin{matrix} & v_0 & \dots & v_v \\ \begin{matrix} w_0 \\ \vdots \\ w_{w'} \end{matrix} & \begin{bmatrix} \cdot & & \\ & \ddots & \\ & & \cdot \end{bmatrix} & \end{matrix} \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

W' is subset of words for which we have brain data

$$A = \begin{matrix} & l_0 & \dots & l_l \\ \begin{matrix} w_0 \\ \vdots \\ w_{w'} \\ w_w \end{matrix} & \begin{bmatrix} \cdot & & \\ & \ddots & \\ & & \cdot \end{bmatrix} & \end{matrix} \begin{bmatrix} \\ \\ \\ \end{bmatrix} \quad \textcolor{red}{A'}$$

$$\operatorname{argmin}_{A, D^{(c)}, D^{(b)}} \sum_{i=1}^{\boxed{w}} \|X_{i,:} - A_{i,:} \times D^{(c)}\|^2 + \sum_{i=1}^{\boxed{w'}} \|Y_{i,:} - A_{i,:} \times D^{(b)}\|^2 + \lambda \|A\|_1$$

$$D^{(b)} = \begin{matrix} & v_0 & \dots & v_v \\ \begin{matrix} l_0 \\ \vdots \\ l_l \end{matrix} & \begin{bmatrix} \cdot & & \\ & \ddots & \\ & & \cdot \end{bmatrix} & \end{matrix} \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$




Advantages of JNNSE

- Handle partially paired data. Vs Canonical Correlation Analysis or Partial Least Squares that require data about the same observations in both cases.
- No need to have a common average brain (can concatenate activation across subjects in the Y matrix, per word)
- Merge different brain imaging experiments adding a specific loss

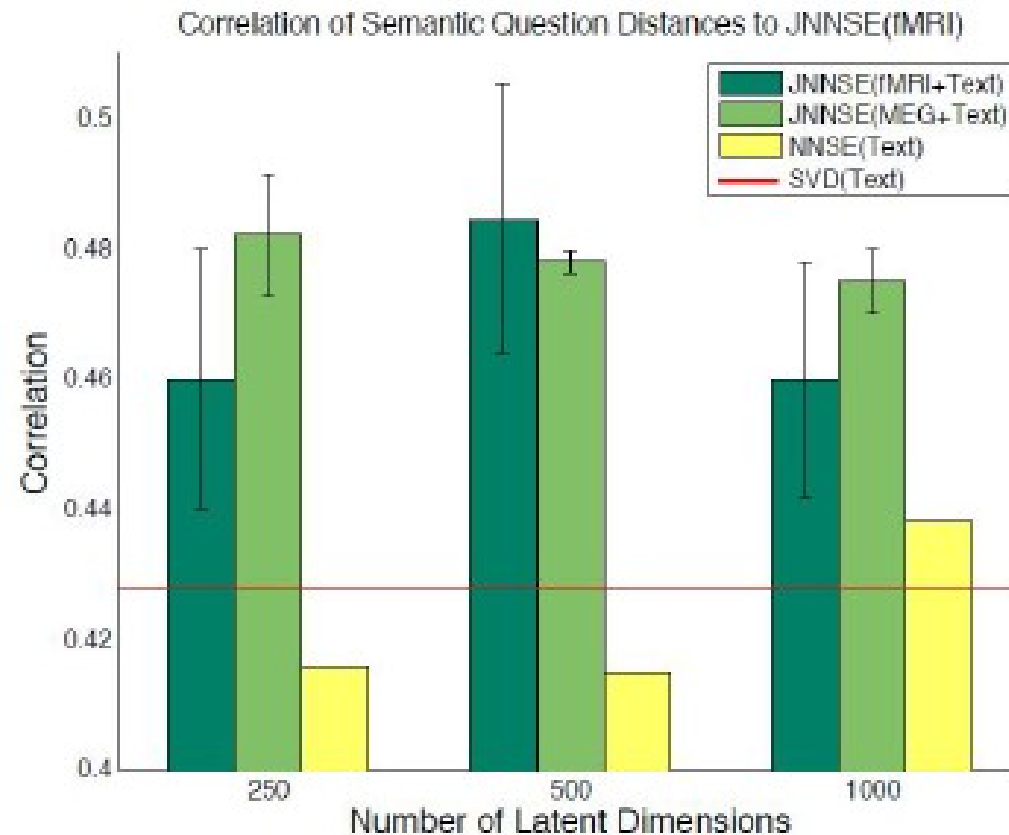


Experiments

Prediction of semantic annotations.

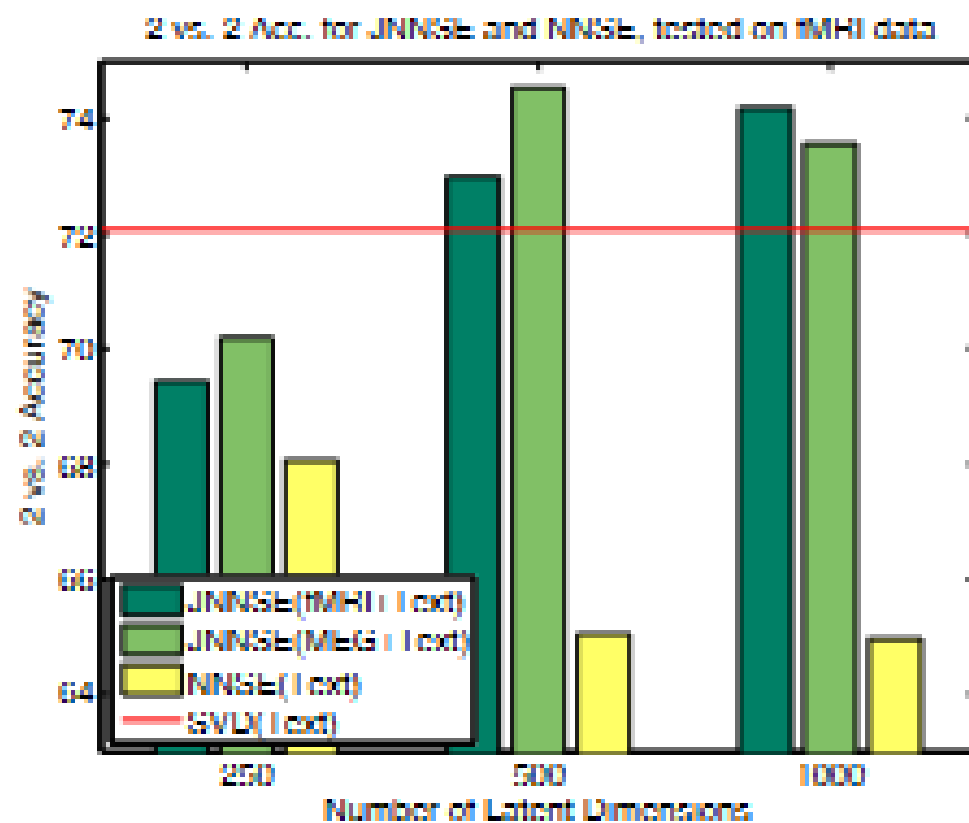
- Step1: obtain behavioral measure of semantics
- Step2: try predicting the behavioral measures using VSMsa
- Step1: 60 words; each with 218 semantic annotations.
- Step2: learn to predict the 60x218 matrix from 60x**A** matrix obtained from
 - JNNSE(Brain+Text); [using 250, 500 or 1000 latent D]or
 - NNSE(text alone).
 - Note: these create different A matrices.
- Assessment: pairwise (Euclidean) distances in 218D space (JNNSE vs. Human Behavior)

Correlation with behavioral data



Word prediction from brain data.

- Matrix A from JNNSE or NNSE used as outcome variable
 - They will predict it (for left out words) one column at a time.
- For each lower-dimension, there is a set of values over words (Y)
- Regression is used to predict the value of that dimension l over the Y words.
 - This is repeated for all l dimensions..
 - ..which produces a predictive l -dim vector per word.
- They train the regression on Matrix A (JNNSE / NNSE) consisting of 58 words. They then produce predictive vectors for 2 left out words and evaluate their similarity to the ground truth of those embeddings in A .

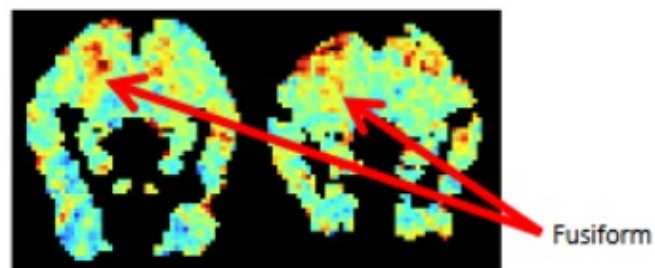


Prediction of corpus data

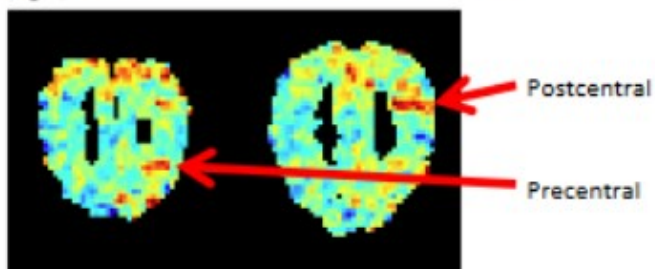
- Context of application: produce X-matrix (“raw”) entries for words for which there is not enough corpus data required for creation of corpus statistics. That is, predict actual corpus data.
- Principle, use JNNSE to obtain A-entries for words appearing in X and Y (corpus and brain), but also some words appearing only in Y (brain).
- Using the $D(c)$ in Eq 4, recreate the entries for those words.
- They compute a rank accuracy measure (I read this as rank-1) and the mean rank accuracy is as high as 67% for $l=500$.

Mapping semantics onto the brain

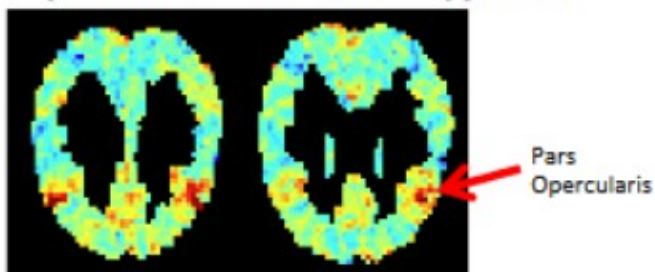
- $D(b)$ (slide 12) is a matrix of l (latent dimension) by v (voxels).
- This allows us to obtain a brain map for each dimension l in that matrix.
- They tweak the importance of the perceptual features (Y) by scaling their values (details missing; but we can consider weighting schemes)
- Information about dimension: strong-loading-words procedure discussed above.
 - They plot these dimensions on brain slices.



(a) $D^{(b)}$ matrix, subject P3, dimension with top words bathroom, balcony, kitchen. MNI coordinates $z=-12$ (left) and $z=-18$ (right). Fusiform is associated with shelter words.



(b) $D^{(b)}$ matrix; subject P1; dimension with top words ankle, elbow, knee. MNI coordinates $z=60$ (left) and $z=54$ (right). Pre- and post-central areas are activated for body part words.



(c) $D^{(b)}$ matrix; subject P1; dimension with top scoring words buffet, brunch, lunch. MNI coordinates $z=30$ (left) and $z=24$ (right). Pars opercularis is believed to be part of the gustatory cortex, which responds to food related words.

Figure 4: The mappings ($D^{(b)}$) from latent semantic space (A) to brain space (Y) for fMRI and words from three semantic categories. Shown are representations of the fMRI slices such that the back of the head is at the top of the image, the front of the head is at the bottom.

Conclusions

- VSM can be extended or even substituted using brain data
- Addition of brain data strongly improves prediction of human annotations (perhaps substitute?)
- Addition of brain data strongly improves prediction of latent-dimension scores produced in the joint embedding (noise reduction?)
- It is possible to use brain data to synthesize raw corpus data for those words
- Solutions of joint embeddings can be mapped onto the brain space
- Potential modifications: relative weighting of importance in reconstructing X and Y matrices.