A Latent Feature Analysis of the Neural Representation of Conceptual Knowledge

Kai-min Chang, Brian Murphy, and Marcel Just

Carnegie Mellon University, Pittsburgh PA 15206, USA

Abstract. Bayesian probabilistic analysis offers a new approach to characterize semantic representations by inferring the most likely feature structure directly from the patterns of brain activity. In this study, infinite latent feature models [1] are used to recover the semantic features that give rise to the brain activation vectors when people think about properties associated with 60 concrete concepts. The semantic features recovered by ILFM are consistent with the human ratings of the shelter, manipulation, and eating factors that were recovered by a previous factor analysis. Furthermore, different areas of the brain encode different perceptual and conceptual features. This neurally-inspired semantic representation is consistent with some existing conjectures regarding the role of different brain areas in processing different semantic and perceptual properties.

1 Introduction

Mitchell et al. [2] showed that word features computed from the occurrences of stimulus words (within a trillion-token Google text corpus that captures the typical use of words in English text) can predict the brain activity associated with the meaning of these words. The advantage of using word co-occurrence data is that semantic features can be computed for any word in the corpus - in principle any word in existence, as opposed to manually defined semantic features like [3], [4]. Nonetheless, despite the success of this model, the work leaves open the question about how to determine the optimal set of semantic features. [2] hand-picked a set of semantic features defined by 25 verbs: *see, hear, listen, taste, smell, eat, touch, rub, lift, manipulate, run, push, fill, move, ride, say, fear, open, approach, near, enter, drive, wear, break,* and *clean*. This selection process was motivated by conjectures regarding the centrality of sensory-motor features in neural representations of concepts [5]. However, it is likely that there are other sets of semantic features that better characterize the brain activity. One could exhaustively search for the optimal set of features, but such an approach would be computationally intractable and certainly not a satisfying approach.

In this study, we address the question by taking a bottom-up approach. Instead of searching for the optimal set of features that can account for the brain activity, we try to infer the most likely feature structure directly from the patterns of brain activity. We take a generative approach and model the semantic representation as some hidden variables in the probabilistic Bayesian framework.

G. Langs et al. (Eds.): MLINI 2011, LNAI 7263, pp. 234–241, 2012.

A generative process is used to describe how brain activity is generated from this latent semantic representation. The basic proposition is that the human semantic knowledge system is capable of producing an unbounded number of features associated with a concept; however, only a subset of them are actively recalled and reflected in brain activity during any given task. Moreover, some features will be shared among a group of concepts (e.g. both dogs and cows *have four legs*), while some features will be idiosyncratic to particular concepts (e.g. *produces milk* is highly salient for cows only).

Thus, a set of latent indicator variables is introduced to indicate whether a feature is actively recalled. By describing the prior distribution of these latent indicator variables and the distribution of the observed brain activity given the assignment of these latent variables, standard Bayesian inference procedure can be used to infer the recalled features. More specifically, we used the infinite latent feature model (ILFM) with an Indian Buffet Process (IBP) prior [1] to derive a binary feature representation of conceptual knowledge from the brain activity. ILFM is especially suited for our task because it automatically determines the number of features that are manifested in the data. This data-driven feature representation is neurologically-informed and may better capture what people were thinking. To foreshadow our results, the ILFM is able to capture a latent semantic representation that is consistent with human ratings of three semantic factors recovered by factor analysis. Furthermore, we show that the recovered latent features are consistent with some existing conjectures regarding the role of different brain areas in processing different psycholinguistics features.

In section 2, we describe the data set and how areas of interests are identified. In particular, we show that the distributed pattern of brain activity contains sufficient signal to discriminate among concepts. In section 3, we discuss the infinite latent feature model and show how it can be used to recover the latent semantic representation encoded by brain activity. In section 4, we try to interpret the recovered latent features by correlating the latent features with the human ratings of the shelter, manipulation, and eating factors, as well as some psycholinguistic word features. Finally, we discuss some of the implications of our work and suggest some future studies.

2 Experimental Paradigm and Identifying Areas of Interest

We used the CMU fMRI data-set of nine English speakers (5 female, all right-handed, age between 18 and 32) thinking about 60 concrete concepts, in 12 categories, which was previously collected and made available online by [2]. For each concept there are 6 instances of $\sim$20k neural activity features (brain blood oxygenation levels). In an concept-contemplation task, participants were presented with 60 line drawings of concepts with text labels for 3s (followed by a 7s rest period) and were instructed to think of the same properties of the stimulus concept consistently during each presentation.

Table 1. Classification and infinite latent feature analysis

METRIC	ALL	FRONTAL	TEMPORAL	PARIETAL	OCCIPITAL
Rank Accuracy	0.81	0.58	0.70	0.66	0.80
R^2	0.77	0.66	0.69	0.69	0.76
K_+	14.44 ± 3.09	16.67 ± 4.47	14.22 ± 3.67	15.44 ± 6.13	14.89 ± 4.81

Before progressing to the main results, we first attempt to verify if the distribution of brain activity encodes sufficient signal to decode the mental state associated with viewing and contemplating particular concepts. Given the evoked patterns of brain activity (mean PSC) brain that were observed while participants contemplated one of the 60 presented concepts, Gaussian Naive Bayes classifiers were trained to identify the associated cognitive state. For instance, the classifier should predict which of the 60 exemplars the participant was viewing and thinking about.

Classification results were evaluated using 6-fold cross validation, where one of the 6 repetitions was left out for each fold. The voxel selection procedure was performed separately inside each fold, using only the training data. Since multiple classes were involved, rank accuracy was used as an evaluation metric, as in [2]: given a new fMRI image to classify, the classifier outputs a rank-ordered list of possible class labels from most to least likely. The rank accuracy is defined as the percentile rank of the correct class in this ordered output list, ranging from 0 to 1. Classification analysis was performed separately for each participant, and the mean rank accuracy was then computed over the participants.

The first row in Table 1 shows the results of the classification analysis. All classification accuracies were significantly higher than chance ($p < 0.05$), where the chance level for each classification is determined based on the empirical distribution of rank accuracies over 100 randomly permuted null models. Using activities recorded throughout the brain, the classifier was able to distinguish among the 60 exemplars with mean rank accuracies close to 81%. Distinct classifiers were also trained separately for several anatomical regions: the frontal, temporal, parietal, and occipital lobes. Occipital lobe activity gives the best classification accuracies, but the temporal, parietal, and frontal lobes can also classify with accuracies significantly higher than chance. High classification accuracies indicate that the distributed pattern of brain activity does encode sufficient signal to discriminate differences among stimuli. Knowing this, we can turn to the question of what semantic representation is encoded in brain activity.

3 Learning a Semantic Representation from Brain Activity

We used the infinite latent feature model (ILFM) with an Indian Buffet Process (IBP) prior [1] to derive a binary feature representation of conceptual knowledge

from the brain activity. [1] described a non-parametric Bayesian approach to latent variable modeling in which the number of variables is unbounded.

Let X denote the brain activity recorded in our concept-contemplating task and Z denote the latent semantic representation that underlies the brain activity pattern. The infinite latent feature model is then specified by 1) a prior over the feature vectors $P(Z)$, and 2) a distribution over the brain activity matrices conditioned on the feature assignments, $p(X|Z)$. In a linear-Gaussian infinite latent feature model, the distribution of Z is modeled with an IBP prior, and the distribution of $X|Z$ is assumed to be matrix Gaussian with mean ZA and variance $\sigma_X I$. The following equations summarize the linear-Gaussian infinite latent feature model.

$$Z \sim \text{IBP}(\alpha, \beta) \tag{1}$$

$$A \sim \text{Gaussian}(0, \sigma_A^2 I) \tag{2}$$

$$X|Z, A, \sigma_X \sim \text{Gaussian}(ZA, \sigma_X^2 I) \tag{3}$$

In the context of the 60-words experiment, X is a matrix of size $N \times V$, where x_{nv} is the brain activity for concept n at voxel v. $N = 60$ and $V = 120$ since our stimulus set consists of 60 concepts and a voxel selection procedure used in [2] identified the 120 most stable voxels. Notice that each concept was presented 6 times in our experiment; a representative fMRI image for each concept was created by computing the mean fMRI response over the 6 presentations, and the mean of all 60 of these representative images was then subtracted from each brain activity vector.

Z is a matrix of size $N \times K$, where z_{nk} is a binary value indicating if the feature k is recalled for concept n. By assuming an IBP prior on the distribution of Z, the number of K is unbounded. The hyper-parameters α and β controls the number of features per concept and the total number of features in the matrix, respectively.

A is matrix of size $K \times V$, where a_{kv} denote the feature-to-activity mapping, such that $X = Z \times A$. By assuming that the distribution of A is matrix Gaussian with mean 0 and variance $\sigma_A I$, we can easily integrate out A when computing the full distribution of $P(Z) \cdot p(X|Z)$.

We used Gibbs Sampling [6] to infer Z. The Gibbs sampler was initialized with $K_+ = 1$, with a random assignment to the first column by setting $z_{i1} = 1$ with probability 0.5. The model parameters, α, β, σ_A, and σ_X were all initially set to 0.5, and then sampled by adding Metropolis-Hastings [7] steps to the MCMC algorithm. Separate ILFM is estimated for each participant and each brain region. The sampler was allowed to run for 1000 iterations (though it typically converged after approximately 100 iterations). Rows 2 and 3 in Table 1 show the amount of systematic variance (R^2) accounted by the latent semantic structure and the average number of latent features (K_+) inferred from the brain activity in each brain region. All R^2 were significantly higher ($p < 0.05$) than chance, where the chance level of approximately 0.23 was determined by random assignments to the latent semantic matrix.

4 Interpreting the Latent Features

The question now is what does each latent feature mean? Do different brain areas encode different types of conceptual features? We can try to find correlations between each latent feature vector (column vector describing which concepts possess this feature) and both semantic features of the concepts (human ratings of the 60 concepts reported in [8]), and possible psycholinguistic confounds (taken from the MRC Psycholinguistic Database [9]).

4.1 Independent Human Rating

Just et al. [8] used factor analysis to identify three semantic factors: manipulation, eating, and shelter that provide a good basis for the representation of the 60 concepts. The manipulation factor assigns high scores to concepts that are held and manipulated with one's hands (e.g. pliers, screwdriver). The eating factor assigns high scores to concepts that are edible (e.g. vegetables) or are instruments for eating or drinking (e.g. glass, cup). The shelter factor assigns high scores to concepts that provide shelter (e.g. house, apartment) or entry to a sheltering enclosure (e.g. airplane). They collected an independent set of ratings of each word with respect to each of the three semantic factors from a separate set of 14 participants. For example, for the eating-related factor, participants were asked to rate each word on a scale from 1 (completely unrelated to eating) to 7 (very strongly related).

We show that the latent features recovered by ILFM are consistent with the human ratings of the shelter, manipulation, and eating factors that are recovered by the factor analysis. For each latent feature inferred, we correlate the latent feature vector (column vector describing which objects possess this feature) with human ratings of the three semantic factors (column vector describing how human rate the relatedness between the 60 objects and the specified factor). For each brain region, we identify the maximum correlation between the semantic factors with any one of the latent semantic feature. Figure 2 shows the maximum correlation between the latent feature vector and human rating vector, averaged across subjects. The error bars indicate 95% confidence intervals, where the distribution of that statistic is estimated from the 900 Gibbs samples (excluding the first 100 burn-in samples). Notice that the magnitude of correlations are low partly because we are correlating binary latent feature vectors against semantic and psycholinguistic features that are continuous.

Different brain regions are biased toward different latent features: the frontal lobes tend to infer latent features that correlate with human ratings of manipulation, whereas the temporal and parietal lobes tend to infer latent features that correlate with human ratings of shelter and eating factor, respectively. This pattern of results is consistent with contemporary conjectures that the pre-central area in the frontal lobe is involved with motor planning, the fusiform and parahippocampal place areas that are included in our temporal lobe are involved with thought about places, and parietal area is involved in aggregation of sensory input.

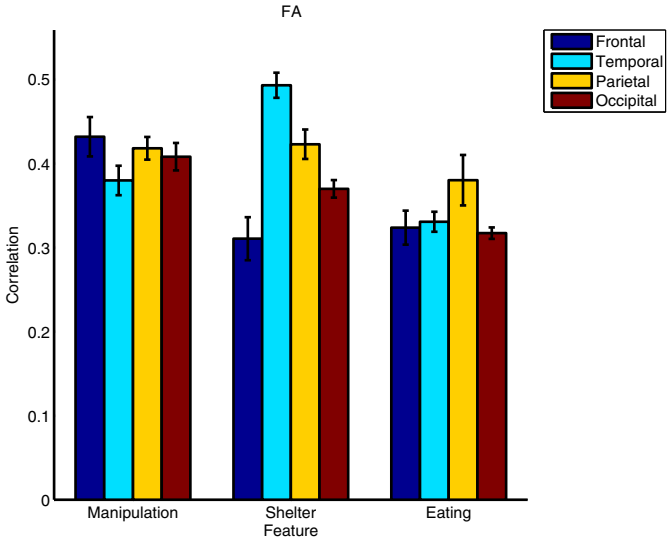

Fig. 1. Correlating the latent features with human ratings of shelter, manipulation, and eating factor

4.2 MRC Psycholinguistic Database

The MRC Psycholinguistic Database [9] is a dictionary that contains 150837 words with up to 26 attributes for each word which are relevant to linguistic processing. While lexical measures are defined for most of the words, psychological measures are recorded for only about 2500 words. Some of the psycholinguistic measures that are of interest to us include meaningfulness (cmean), familiarity (fam), concreteness (cnc), imaginability (img), number of letters (nlet), number of phonemes (nphn), and frequency (t-lfrq).

For each latent feature inferred, we also correlate the latent feature vector (column vector describing which objects possess this feature) with each of the MRC psycholinguistic measure (column vector describing the psycholinguistic score of the 60 objects). Figure 2 shows the maximum correlation between the latent feature vector and MRC feature vector, averaged across subjects. Again, different brain regions infer different latent features: frontal lobe activity correlates most with meaningfulness, although the correlation is not significantly different from that of the temporal and parietal lobe. The parietal lobe shows a bias for concreteness and imaginability, compared to the other brain regions. The temporal lobe tends to encode features that correlate with number of phonemes in a word, consistent with the existing conjecture that the temporal lobe is involved in speech production. Notice that the occipital lobe tends to encode features that correlate most strongly with the number of letters, but not the number of phonemes.

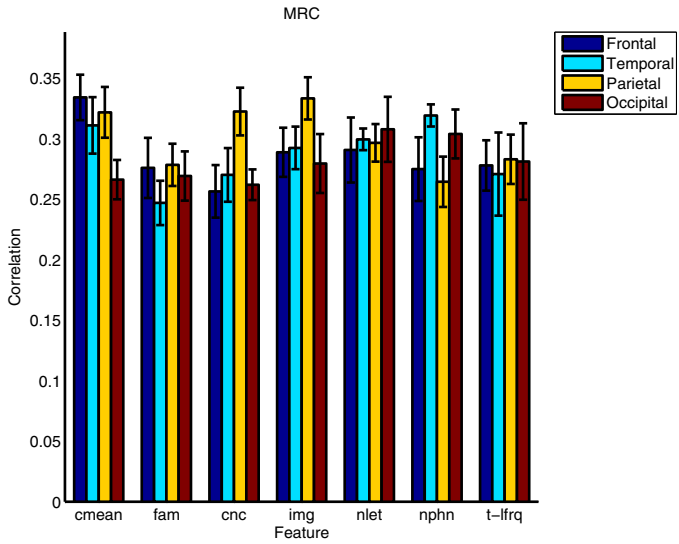

Fig. 2. Correlating the latent features with MRC psycholinguistics features.

5 Conclusions and Future Directions

In this study we use a generative probabilistic model to describe how fMRI-measured brain activity reflects a latent semantic representation. This data-driven feature representation is neurologically-informed and may better capture what people were thinking.

Compared to factor analysis (FA) or multi-dimensional scaling (MDS), there are several advantages of using ILFM to model the semantic representation that underlie brain activity, which 1) offers a formal probabilistic account of the brain activity, 2) automatically determines the number of features that are manifested in the data, and 3) allows different number of features to be inferred per words. One critical difference between ILFM and FA/MDS is that the latter use a continuous representation. In this study, we use a binary representation of the feature matrices, but it can be easily extended to a continuous representation. [1] showed that the binary matrix Z can be combined with a continuous matrix V to define a richer representation.

There are several possible extensions of this work. First, in this study we try to interpret the learned latent semantic features by comparing the vectors to human ratings of three semantic factors and MRC psycholinguistic word features, but one shouldn't stop here. One obvious direction is to compare the feature vector with other types of lexical semantic feature, such as elicited property lists [4] and word co-occurrence statistics [10]. Moreover, we inferred the latent features from predetermined brain regions that are known to process certain semantic and psycholinguistics features, such that we can demonstrate that ILFM can be

used to verify some existing conjectures. An extension is to infer latent features from brain regions whose processing role are unknown in an attempt to discover new areas of interest. Finally, in this work we fitted a unique model for subject, it is interesting to explore how ILFM scale up to incorporate multiple subjects and discover feature representation that generalizes across people.

Acknowledgments This research was supported by the National Science Foundation, Grant No. IIS-0835797, and by the W. M. Keck Foundation. We would like to thank Jennifer Moore for help in preparation of the manuscript.

References

1. Griffiths, T.L., Ghahramani, Z.: The Indian Buffet Process: An Introduction and Review. Journal of Machine Learning Research 12, 1185–1224 (2011)
2. Mitchell, T., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A.: Predicting human brain activity associated with the meanings of nouns. Science 320, 1191–1195 (2008)
3. Kipper, K., Dang, H.T., Palmer, M.: Class-based construction of a verb lexicon. In: Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, Austin, Texas, pp. 691–696 (2000)
4. Cree, G.S., McRae, K.: Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). Journal of Experimental Psychology: General 132(2), 163–201 (2003)
5. Caramazza, A., Shelton, J.R.: Domain-specific knowledge systems in the brain the animate inanimate distinction. Journal of Cognitive Neuroscience 10(1), 1–34 (1998)
6. Geman, S., Geman, D.: Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(6), 721–741 (1984)
7. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics 21(6), 1087–1092 (1953)
8. Just, M.A., Cherkassky, V.L., Aryal, S., Mitchell, T.M.: A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS ONE 5, e8622 (2010)
9. Coltheart, M.: The MRC Psycholinguistic Database. Quarterly Journal of Experimental Psychology 33A, 497–505 (1981)
10. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16, 22–29 (1990)