Predicting human perceived similarity across a wide range of object categories via sparse positive embeddings

Martin N Hebart (martin.hebart@nih.gov)
Laboratory of Brain & Cognition, National Institute of Mental Health
10 Center Drive, Bethesda, MD 20814 USA

Charles Zheng (charles.zheng@nih.gov)
Section on Functional Imaging Methods, National Institute of Mental Health
10 Center Drive, Bethesda, MD 20814 USA

Francisco Pereira (francisco.pereira@nih.gov)
Section on Functional Imaging Methods, National Institute of Mental Health
10 Center Drive, Bethesda, MD 20814 USA

Chris I Baker (bakerchris@mail.nih.gov)
Laboratory of Brain & Cognition, National Institute of Mental Health
10 Center Drive, Bethesda, MD 20814 USA

Abstract:

How do humans represent behaviorally-relevant dimensions of real-world objects? To address this question, we recently used a triplet odd-one-out task to collect >800,000 behavioral judgments on images of 1,854 diverse basic-level object categories. To explain human behavior and characterize the similarity between pairs of objects, we developed a simple cognitive model that yielded sparse, interpretable perceptual and conceptual dimensions. To determine the utility of those dimensions, here we investigate two questions. First, to what degree can we predict those dimensions from a semantic embedding (Pilehvar & Collier, 2016) and activations in a deep convolutional neural network (CNN)? Second, can we use those predicted dimensions to reconstruct human behavioral similarity? To address these questions, we applied Ridge and Elastic Net regression to semantic embeddings and the activations in fully-connected layer 7 of the CNN VGG-16, respectively. We related the performance to two baseline models: The computational models alone, and a recently proposed method that transforms model features (Peterson et al., 2016). Our results demonstrate excellent prediction of many dimensions and strongly improved predictions of behavioral similarity using our model as compared to both baseline models. These results represent an important step towards both predictive and interpretable models of human cognitive representations.

Keywords: object recognition; similarity; prediction

Introduction

Humans can categorize objects according to an almost infinite number of criteria, yet some dimensions of objects matter more for our everyday behavior than others. For example, a fir tree may be characterized by its specific shape, by being a natural object, by having needles, or by its utility as a Christmas tree; however, the use of its oil in perfumes may be seen as a less relevant dimension. One view, then, is that our mental representations of objects are determined by the behaviorally-relevant object dimensions (Kourtzi & Connor, 2011).

A central approach for inferring these representations is to measure the perceived similarity of objects (Shepard, 1987). However, given the large number of existing object categories and the wide range of visual appearances, it is challenging to reveal principles of object representations that would generalize to a wider range of categories or object stimuli in general.

The fields of computer vision and natural language processing recently have seen strong advances through the development of deep convolutional neural networks (CNNs) that rival human performance (e.g. Simonyan & Zisserman, 2015), and the development of semantic embeddings such as word2vec (Mikolov et al., 2013) that yield strongly improved performance over previous methods in extracting feature vector representations for words from documents. Both of these methods produce representations of objects, when given as input an image or a word naming the object. They may, therefore, contain enough information to allow the prediction of human perceived similarity.

Recently, Peterson et al. (2016) demonstrated that, after simple reweighting, the similarity of object representations within CNNs exhibits a strong correspondence to perceived similarity within several object categories (e.g. animals), although with limited generalization between categories. For semantic embeddings, much less is known about their relationship between object representations and behavioral similarity, although direct correlations have revealed only a weak correspondence (Bankson et al., 2018).

Given that humans may base their similarity ratings on the behaviorally-relevant dimensions of objects, one approach to predicting behavioral similarity might be to first use CNNs and semantic embeddings to predict the dimensions that those similarity ratings are based on, and then to reconstruct behavioral similarity from those dimensions.

The goal of the present work is twofold. First, we aim at verifying the utility of a set of sparse, interpretable dimensions extracted from a large number of behavioral similarity ratings across 1,854 object categories, by predicting them from CNNs and semantic embeddings. Second, we aim at testing the usefulness of CNNs and semantic embeddings in predicting human behavioral similarity for a wide range of object categories, with the goal of offering a model that can be used to accurately predict behavioral similarity for a wide range of object categories.

Methods

Behavioral responses were collected for images of 1,854 basic-level object categories, with one representative image per category. These basic-level objects were selected to be representative of the English language (for details, see Dickter et al., 2018). dimensions.

We used Amazon Mechanical Turk to collect 831,960 trials of a triplet odd-one-out task, in which three object images are presented side-by-side and a worker has to choose which object is the most dissimilar. This task measures the perceived similarity of those objects, but has the advantage of providing bias-free responses and allows measuring similarity across a wide range of contexts imposed by the other objects in the triplet.

Since it was not possible to collect all similarity ratings across all contexts (> 100 billion combinations), we developed a simple cognitive model based on sparse positive embeddings that is optimized for predicting behavioral choices while concurrently providing similarity ratings and extracting the behavioral similarity of objects. This approach provided us with a set of 36 interpretable perceptual and conceptual object dimensions.

In the next step, we sought to test to what degree we can predict those dimensions from semantic embeddings and activations in a CNN. As a semantic embedding, we chose a sense representation based on de-conflated word representations (Pilehvar & Collier, 2016), which provides more interpretable dimensions for each object by using the semantic structure provided through synsets in WordNet (Fellbaum, 1998). Sense vectors were available for 1,796 object categories. As a CNN, we chose the fully-connected layer 7 in VGG-16 (Simonyan & Zisserman, 2015), which has previously been shown to provide good correspondence to behavioral similarity (Peterson et al., 2016). The sense vectors had 300 dimensions, while the CNN layer had 4,096 dimensions.

The predictions of the 36 dimensions based on the sense vectors were carried out using Ridge regression, while the predictions based on CNN activations were carried out using Elastic Net regression. We used 10-fold cross-validation for predictions, and nested 10-fold cross-validation for selection of hyperparameters (for Ridge regression: lambda, for Elastic Net regression: lambda and alpha).

In the next step, we calculated a large-scale similarity matrix by computing the dot product of those 36 predicted dimensions and compared them to the true behavioral similarity.

Finally, as a baseline, we calculated large-scale similarity matrices based only on the sense vectors and the activations in CNN layer 7. In addition, we compared our results to a previously published method (Peterson et al., 2016), which computes a weight for each of the features in a model that determines the similarity.

Results

Using sense vectors and CNN layer 7, many of the 36 dimensions were predicted highly accurately (sense vectors: max R^2: 0.85, median R^2: 0.38; CNN layer 7: max R^2: 0.72, median R^2: 0.39). This demonstrates that representations in semantic embeddings and CNNs carry information highly predictive of many of these dimensions, demonstrating the usefulness of those computational models for predicting behaviorally-relevant dimensions. In addition, the results suggest that the dimensions extracted from our cognitive model are useful for characterizing components of behavioral similarity.

The predicted behavioral similarity yielded a similar pattern of results. Behavioral similarity predicted from sense vectors accounted for 57.9% of the variance, while behavioral similarity predicted from CNN layer 7 explained 49.1%. In contrast, the baseline model based on the sense vectors and CNN layer 7 alone accounted for 26.7% and 18.6% of the variance, while the baseline model based on simple transformations accounted for 43.3% and 41.0%, respectively.

Discussion and Conclusion

Our results demonstrate that it is possible to predict many dimensions of a cognitive model used to characterize behavioral similarity ratings. Since these dimensions are well-predicted and interpretable, this result opens the avenue for interpretable information about the nature of the representations in semantic embeddings and CNNs and how they relate to human behavior.

Further, the fact that more than half of the variance of behavioral similarity ratings for a very wide range of object categories can be predicted from semantic embeddings and CNNs may represent an important step towards a generative model of behavioral similarity.

Previous work using a smaller set of object categories has demonstrated higher prediction accuracy using simple transformations of model features (Peterson et al., 2016), while using the same method in the present study turned out to predict not as well. However, many of those predictions were done within category, and the previous study left open whether those results would generalize to the much larger range of categories used in the present experiment. Alternatively, due to the partial sampling of the matrix it is possible that the estimation of behavioral similarity, and therefore the prediction of behavioral similarity from computational models, can still be improved. We will address this question by improving the sampling of the behavioral similarity matrix.

Finally, another avenue for future work is to combine multiple computational models and test whether they can jointly explain more variance than any model alone.

Together, using a simple cognitive model of behavioral similarity, the results shed light on the utility of computational models for predicting behavioral similarity and open an avenue for the use of interpretable dimensions to reveal the structure of human cognitive representations.

Acknowledgements

This work was supported by the Intramural Research Program of the National Institute of Mental Health (ZIA-MH-002909) and a Feodor-Lynen Fellowship of the Humboldt Foundation awarded to M.N.H.

References

Bankson, B. B., Hebart, M. N., Groen, I. I., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *Neuroimage*.

Dickter, A., Hebart, M. N., Kidder, A., Kwok, W., & Baker, C. I. (2018). A large-scale object database based on comprehensive sampling of the English language. *Poster presented at the Annual Meeting of the Vision Science Society.*

Fellbaum, C. (1998). WordNet: An Electronic Database. MIT Press, Cambridge, MA.

Kourtzi, Z., & Connor, C. E. (2011). Neural representations for object perception: structure, category, and adaptive coding. Annual review of neuroscience, 34, 45-67.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Peterson, J., Abbott, J., & Griffiths, T. (2016). Adapting deep network features to capture psychological representations. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Austin, TX.

Pilehvar M. T., & Collier, N. (2016). De-conflated semantic representations. In: *Proceedings of EMNLP*, Austin, TX.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.