Preserved Structure Across Vector Space Representations

Andrei Amatuni
andrei.amatuni@duke.edu
Department of Psychology
Duke University

Elika Bergelson
elika.bergelson@duke.edu
Department of Psychology
Duke University

Abstract

We find evidence of preserved structure between vector space representations of words and their corresponding image embeddings. This is evidence of regularity between the representations learned using distributional statistics of words and the visual characteristics of those same items. We find that some classes of objects, namely inanimate ones, preserve their within-class structure across these two spaces more strongly than others (e.g. animate objects), and that this quality of preserving class-level relationships across representational spaces might aid in lexical acquisition, with invariance serving as an informative marker of category boundaries. Our current analysis does not show significant AoA benefits for inanimate objects, but does exhibit a stable pattern suggesting that other partitioning schemes might be worth exploring (or something like that)

Keywords: vector space models; semantic similarity; word learning

Introduction

Methods

We generate two sets of vector representations for a common set of words first learned by most infants. The first set of vectors are taken from a pretrained set of GloVe representations (Pennington, Socher, & Manning, 2014), a modern distributional semantic vector space model. The second set is taken from the final layer activations of a pretrained image recognition model, Google's Inception V3 convolutional neural network (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). Both of these representations are what's refered to as "embeddings". They map objects from one medium (e.g. images or words) into a metric space, where distances between points can be computed and function as a measure of similarity between objects.

In the case of our word vectors, the GloVe algorithm instantiates the distributional hypothesis, which proposes that words which co-occur with each other share similar meaning (Firth, 1957; Harris, 1954), and by capturing the covariance of tokens in large text corpora, you capture some aspect of their semantic structure. The image embeddings, on the other hand, are taken from the final layer of activations in a convolutional neural network, whose objective function tunes network parameters in service of object recognition, where the loss function is computed in reference to a set of labeled training images (Russakovsky et al., 2015). The final layer of this network encodes the most abstract and integrated visual features, serving as the basis for classification into 1000 different classes.

Defining a prototypical image

In the case of word vectors, each word is assigned a unique point in a common vector space. Different images containing objects of the same type, on the other hand, will have varying vector representations after passing through the layers of a neural network. This presents a problem in comparing the two forms of representation. We must first define the most prototypical (or average) image vector for any given category of object.

Given a set of images S_c containing objects belonging to a single category c (e.g. cat, dog, chair), we define our prototypical vector $\hat{x}_c$ of S_c as the generalized median within a representational space U. This is the vector with minimal sum of distances between it and all the other members of set S_c in U. If x and y are vectors in space U, resulting from images in S_c being passed through a neural network, then

$$\hat{x}_c = \arg\min_{x \in U} \sum_{y \in U} d(x, y)$$

We define our $d(x, y)$ to be the cosine similarity measure:

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}$$

Our $d(x, y)$ is not a metric in the strict sense, but is less susceptible to differences in L^2 norm influencing our measure of similarity, as is the case with the Euclidean distance. These magnitude difference can be the product of frequency effects in the training data, and the cosine similarity corrects for this.

The image inputs we use are all 960x960 images of a single object on a gray background. These images were chosen by virtue of their presence in infants' early linguistic environment, aggregated as part of the SEEDLingS project, which gathered longitudinal audio and video data of infants' home environments (Bergelson, 2016a, 2016b). We arrive at a set of 27 unique words, selected on the basis of having at least 9 unique images with which to determine the most prototypical. The more images we have of any given category, the more robust our measure of category variance in image vector space, resulting more representative category vectors. These are all words found on WordBank (Frank, Braginsky, Yurovsky, & Marchman, 2017), a compilation of the MacArthur-Bates Communicative Development Inventory, which we use as our proxy for age of acquisition. By studying the behavior of these developmentally salient objects, our analysis is able to speak to the statistical structure of those objects which infants will be most readily contending with.

Comparing spaces

After we have our two sets of vectors (i.e. those from word vector space and those from image vector space), we can compare all the pairwise distances between objects, both within a single space and across the two. When comparing across the two spaces, a correlation in pairwise distances implies that inter-object distances have been conserved. For example, if "dog" and "cat" are close together in word space and mutually far apart from "chair" and "table" in that same space, maintaining this relationship for all pairwise distances in the *other* vector space means that the global inter-object structure is preserved across this mapping, despite being in radically different spaces, both in terms of dimensionality (300 for words, and 2048 for images in our case) and by virtue of using completely different algorithms and inputs to establish the vector representations for objects. So while their absolute locations might have been radically transformed, this correlation would be a measure of the *degree of invariance* in their positioning relative to each other.
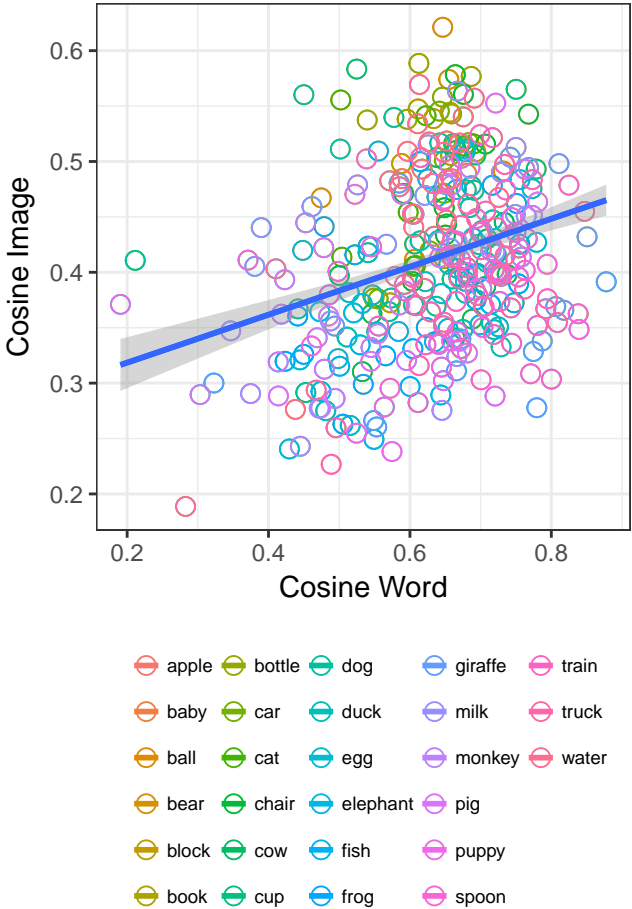
Figure 1: Relative cosine distance between points in word embedding space correlates with relative distance in image embedding space ($R = 0.30$, $p < 9.9e - 16$). Graph contains all pairwise distances for every word.

Results

We find that pairwise cosine distances between objects in word vector space correlate with those same pairwise distances in the image vector space. If we partition the set of inter-word distances into those that are either animate-animate, inanimate-inanimate, or mixed, we find that the pairs of distances between inanimate objects significantly correlate across our two spaces ($R = 0.38$, $p < 1.7e - 07$), while the other two pairings do not.
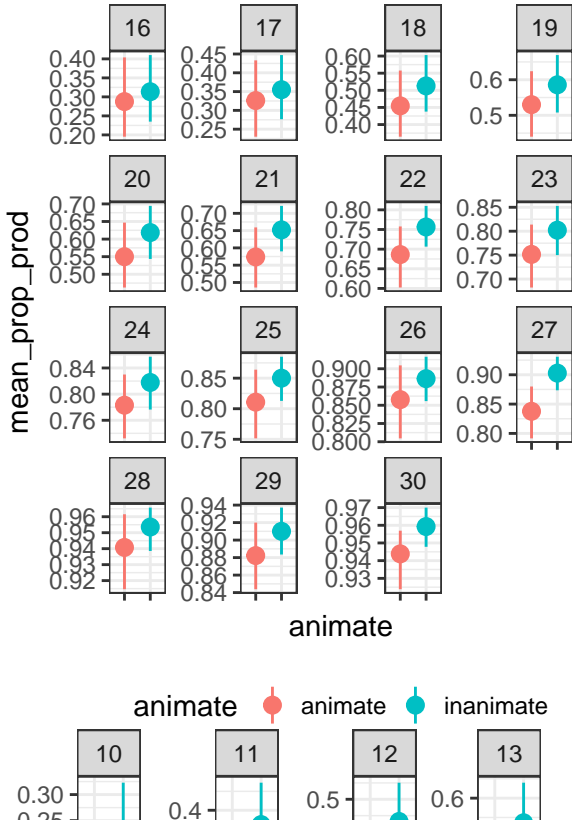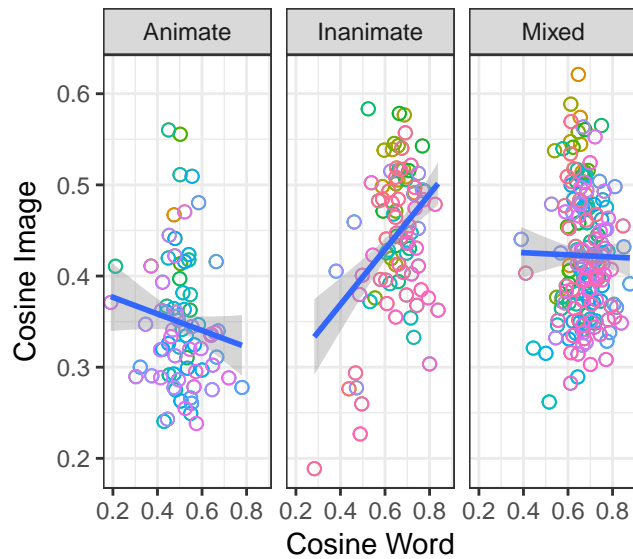
Figure 2: Inanimate objects display a significantly stronger correlation when mapping across vector spaces, meaning that they preserve their within-class structural relationships more reliably across these two spaces. Animate and mixed distances do not correlate. Each graph contains all pairwise distances between objects that are either a) both animate ($R = -0.13$, $p < 0.12$), b) both inanimate ($R = 0.38$, $p < 1.7e - 07$), or c) mixed animate-to-inanimate ($R = -0.01$, $p < 0.8$)
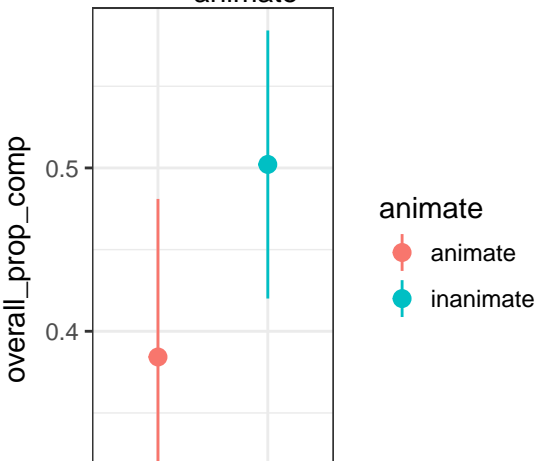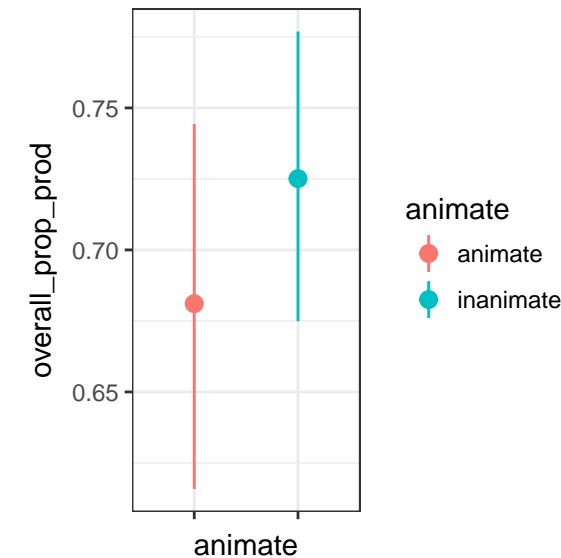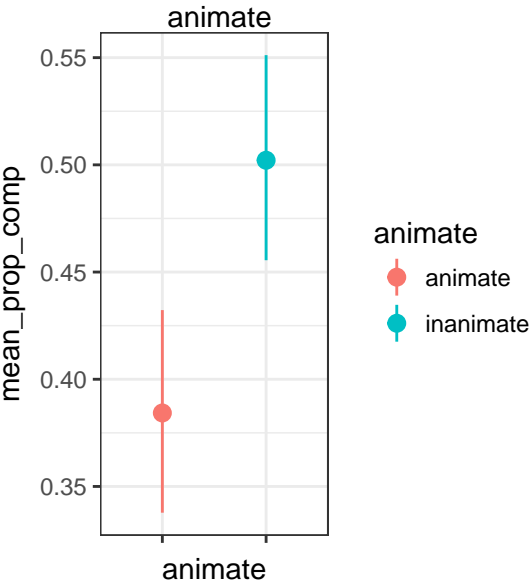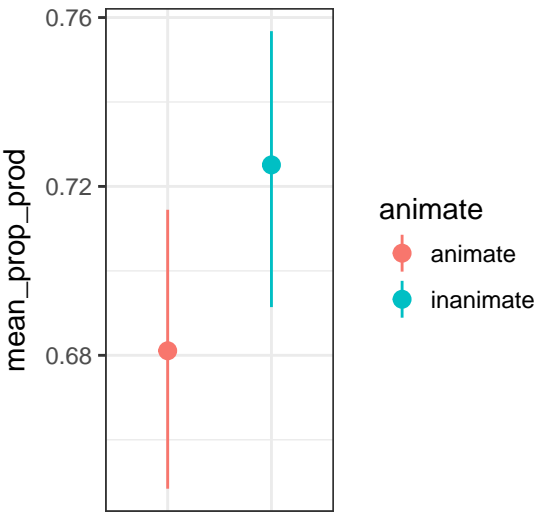
Discussion

Conclusion

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

Bergelson, E. (2016a). Bergelson seedlings homebank corpus. http://doi.org/10.21415/T5PK6D

Bergelson, E. (2016b). SEEDLingS corpus. Retrieved January 26, 2018, from https://nyu.databrary.org/volume/228

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman,

V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183–209.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*(1), 1–32.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1132–1137).

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).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252. `http://doi.org/10.1007/s11263-015-0816-y`

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.