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 age-of-acquisition 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

Infants are presented with a challenge to carve the world into distinct lexical entities in the process of learning their first language. They're provided with little supervision while mapping a territory which William James (1890) dubbed a "great blooming, buzzing confusion". How they determine which aspects of the world to attend to in service of this goal, is an area of ongoing research (Mareschal & Quinn, 2001). Different features of objects and their environments are varyingly informative with regards to object segmentation and category structure. Some researchers have suggested that categorization is along fundamentally perceptual grounds and that only later in development is conceptual knowledge incorporated into these nascent perceptual categories (Quinn & Eimas, 1997, 2000; Quinn, Johnson, Mareschal, Rakison, & Younger, 2000). Others suggest that there are in fact two distinct processes at work, such that perceptual categories are computed automatically by the sensory systems, while conceptual categories are independently formed through conscious action (Mandler, 2000). Träuble and Pauen (2007) provide evidence of functional information (regarding the animacy of objects) influencing early category judgements. Gelman and Markman (1986) explicitly set these two sources of category cues against each other (i.e. functional vs. perceptual), in hopes of discovering which holds greater influence in infant categorization behavior.

The degree to which these two sources of information are separable is an important open question. Any model which hopes to explain the mechanics of human categorization must address how these seemingly disparate forms of information interface in mental representations, and to what degree they interact. In our current study we examine the degree of interaction between representations learned by two different algorithms which operate on apparently dissimilar inputs, namely images and text. These algorithms learn feature representations without hand engineering, purely as a byproduct of their particular training objectives. These training objectives are completely divorced from one another. The features that these algorithms learn are then used to serve their unique practical ends (e.g. machine translation or object recognition in images).

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, products of 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 differences in magnitude between vectors 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 in 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.

Results

We find that pairwise cosine distances between objects in word vector space correlate with those same pairwise distances in the image vector space (see Figure 1). 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 (see Figure 2).

We expect that those classes of objects which preserve their structure between representations more strongly would result in earlier object-referent mappings. This is because inferences about object-referent mappings conditioned on both visual and semantic features would be more stable compared to those cases where the two representations vary independently. For example, an object that is both round (i.e. visual feature) and tends to roll (i.e. semantic feature) would be more salient as a distinct entity than an object whose visual features are entirely uninformative about its functional or semantic qualities.

In our current analysis the class of objects which displays stronger structure preservation (within class) are the inanimate objects. When we partition our set of 27 words into animates and inanimates and plot their relative AoA, we find a noticeable though insignificant preference for inanimates, as expected (see Figures 3 and 4). The choice to partition our set into these two categories is to a degree arbitrary, and we have no reason to believe infants would learn one class of objects earlier than the other. Our current analysis is offered purely as an exploratory exercise, suggesting that perhaps partitions along other taxonomic or associative lines may provide insight in future investigations.

We also examined the degree to which our set of 27 words shared overlapping neighbors in the two vector spaces (see Table 1). We defined a neighbor as all the points with distance less than -1 standard deviation from the mean distance for each word. With this normalized neighborhood threshold, we find that the majority of our words have at least 1 neighbor which is shared across representational spaces.

Discussion

We've reported a significant correspondence between representations learned by two different algorithms operating over seemingly unrelated inputs (i.e. visual and linguistic). What is most noteworthy here is that the only immediate common ground between these representations are the real life objects they both aim to model. This draws us into questions concerning the nature of similarity and the multifaceted character of information which is revealed by objects in the real world. The notion that we can make inferences about one aspect of an object given another aspect, is not surprising or controversial. However, the fact that we can make these bi-directional

word	ratio	neighbors
apple	0.166	**egg**, milk, car, puppy, book, bear
baby	0	puppy, dog, monkey, elephant, frog, cow, pig, train, bear
ball	0	dog, monkey, fish, frog, spoon, cow
bear	0.222	**monkey**, puppy, **elephant**, cow, frog, baby, duck, dog, cat
block	0	egg, water, fish, train, ball, duck, cow, truck, puppy
book	0	car, baby, dog, train, fish, monkey, pig, cat, puppy
bottle	0.333	**water**, **milk**, spoon, egg, baby, cup
car	0.5	**truck**, train
cat	0	dog, puppy, elephant, train, monkey, truck
chair	0	baby, ball, dog, frog, cat, spoon, giraffe
cow	0.222	pig, **elephant**, **monkey**, dog, frog, egg, fish, giraffe, milk,
cup	0.142	**egg**, spoon, milk, duck, fish, water, bottle
dog	0	puppy, cat, cow, frog
duck	0.142	cow, pig, **fish**, egg, frog, elephant, giraffe
egg	0	cow, duck, fish, milk, elephant, spoon, train
elephant	0.25	**cow**, **giraffe**, monkey, frog, train, fish, egg, bear
fish	0.25	**duck**, **pig**, cow, frog, egg, water, monkey, giraffe
frog	0.1	**monkey**, cow, elephant, giraffe, fish, baby, pig, duck, puppy, bear
giraffe	0.428	elephant, **monkey**, **frog**, **cow**, baby, puppy, fish
milk	0.2	bottle, **water**, egg, cow, cup
monkey	0.375	**cow**, **frog**, **pig**, elephant, giraffe, puppy, fish, cat
pig	0.125	cow, **monkey**, duck, puppy, fish, frog, dog, train
puppy	0	dog, pig, monkey, cow, baby, bear, cat, giraffe
spoon	0	egg, frog, cup, ball, fish, milk, giraffe, bottle
train	0.25	**truck**, car, elephant, dog
truck	0.5	car, **train**
water	0.25	bottle, **milk**, fish, egg

Table 1: Overlaps between closest objects in image vector space and word vector space. Neighbors are defined as those other objects which are less than -1 SD from the mean distance for any given word. Those neighbors that are marked red are shared between image and vector spaces. The overlap ratio is the number of shared neighbors across vector spaces divided by the total unique neighbors between the two spaces.

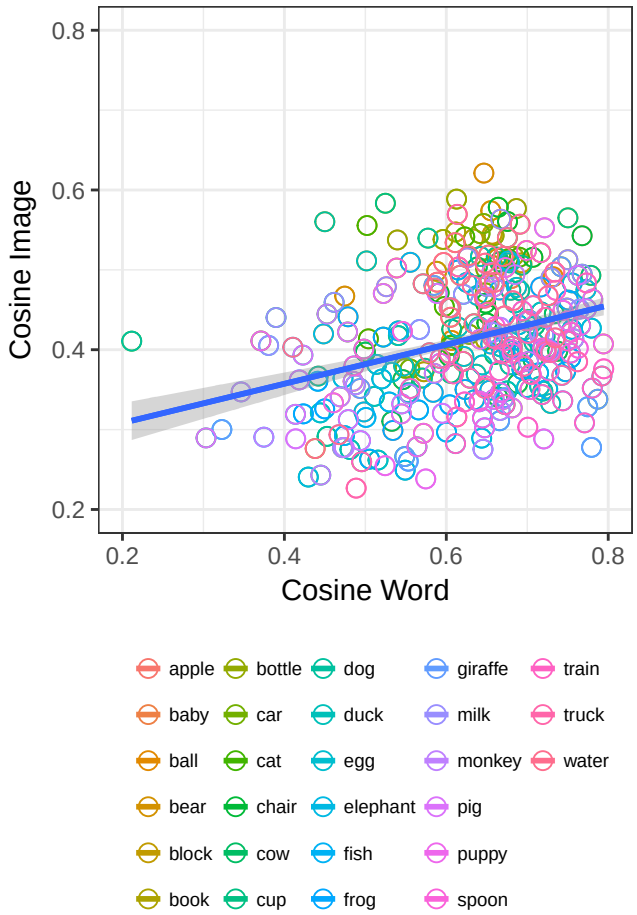

Figure 1: Relative cosine distance between points in word embedding space correlates with relative distance in image embedding space ($R = 0.30$, $p < 9.9e-16$). Graph contains all pairwise distances for every word.

inferences using aspects traditionally treated as being orthogonal, is noteworthy. This is particularly the case given the enormous dimensionality of our feature spaces, and the fact that these algorithms are placed under no pressure to find homologous representations.

Through what metrics can a learning algorithm, or indeed a human, establish gradations of likeness? Are these necessarily the same metrics which form the basis of category boundaries? These are fundamental questions which have enjoyed a long history in the field (Edelman, 1998; Hahn, Chater, & Richardson, 2003; Kemp, Bernstein, & Tenenbaum, 2005; Shepard & Chipman, 1970; Tversky, 1977). While our current work is not sufficient to support a specific mechanism responsible for the observed regularity, it might be indicative of the special role of invariance, given that the unifying thread between our algorithms and inputs are the common objects they represent. Underneath the diversity of visual statistics and token distributions lie stable entities in the world which, by virtue of their invariant actuality, give rise to regularity across measurements at different vantage points (i.e. modali-

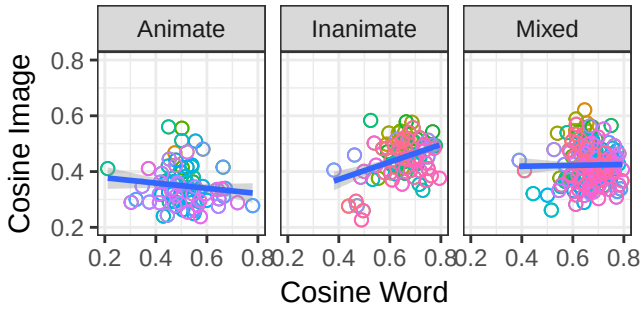

Figure 2: Inanimate objects display a significantly stronger correlation when mapping across vector spaces, meaning that they preserve their within-class structural relationships more reliably across these two spaces. Animate and mixed distances do not correlate. Each graph contains all pairwise distances 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 ($R = -0.01$, $p < 0.8$)

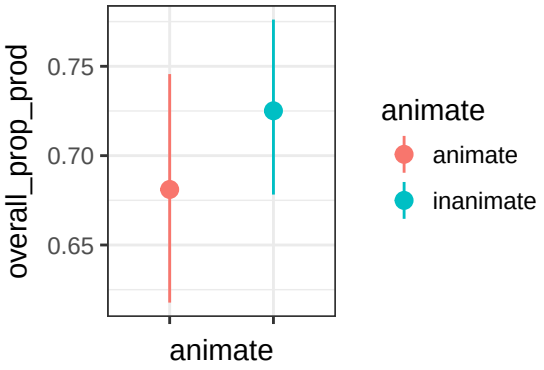

Figure 3: AoA for animates vs inanimates (using child production data) collapsed over month

ties), an idea dating back to Helmhotlz (1878).

We find in our current work that this quality of invariance is differentially present across different classes of entities, namely animate vs. inanimate objects. However, this is conditioned on the particular algorithms we've investigated here, and our extensions into human performance with our AoA anlysis did not show a significant sensitivity to this difference. This could suggest a number things. The first is that humans might not discover the regularities that these algorithms do. Or it could be that our current class partitioning does not provide sufficient contrast in invariance to register human AoA differences. Or it could be that regularity is not a determining factor in ease of acquisition. Of these three, the last is least likely to be the case.

Conclusion

We find evidence of an interaction between visual and semantic features learned by two distinct machine learning algorithms which operate over drastically different inputs, and are

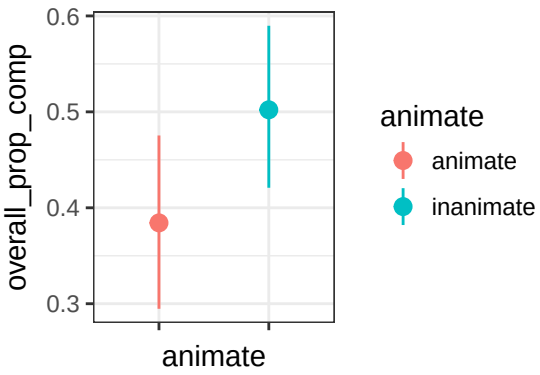

Figure 4: AoA for animates vs inanimates (using child comprehension data) collapsed over month

trained in the service of seemingly unrelated ends. This interaction is indicative of conserved structure between these two supposedly independent sources of information (i.e. visual and functional). If humans are sensitive to this relationship, as these algorithms seem to be, we expect that those classes of object which are more strongly invariant across feature spaces would be more easily learned by infants. We find a noticeable though insignificant relationship between this property and AoA in our current partitioning scheme (animates vs. inanimates).

Acknowledgements

We thank the SEEDLingS team, and NIH DP5-OD019812.

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

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467.

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.

Helmholtz, H. (1878). The facts of perception. *Selected Writings of Hermann Helmholtz*, 1–15.

James, W. (1890). *The principles of psychology*. Henry Holt; Company.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th an-*

nual conference of the cognitive science society (pp. 1132–1137).

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development*, *1*(1), 3–36.

Mareschal, D., & Quinn, P. C. (2001). Categorization in infancy. *Trends in Cognitive Sciences*, *5*(10), 443–450.

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).

Quinn, P. C., & Eimas, P. D. (1997). A reexamination of the perceptual-to-conceptual shift in mental representations. *Review of General Psychology*, *1*(3), 271.

Quinn, P. C., & Eimas, P. D. (2000). The emergence of category representations during infancy: Are separate perceptual and conceptual processes required? *Journal of Cognition and Development*, *1*(1), 55–61.

Quinn, P. C., Johnson, M. H., Mareschal, D., Rakison, D. H., & Younger, B. A. (2000). Understanding early categorization: One process or two? *Infancy*, *1*(1), 111–122.

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

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*(1), 1–17.

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).

Träuble, B., & Pauen, S. (2007). The role of functional information for infant categorization. *Cognition*, *105*(2), 362–379.

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