Mental representations of objects reflect the ways in which we interact with them

Ka Chun Lam[1], **Francisco Pereira**[1], **Maryam Vaziri-Pashkam**[2], **Kristin Woodard**[2], **Emalie McMahon**[3]
[1]Machine Learning Team, [2]Laboratory for Brain and Cognition
National Institute of Mental Health, Bethesda, MD 20892, USA
{kachun.lam, francisco.pereira, maryam.vaziri-pashkam, kristin.woodard}@nih.gov

[3]Department of Cognitive Science
Johns Hopkins University, Baltimore, MD 21218, USA
emaliemcmahon@jhu.edu

Abstract

In order to interact with objects in our environment, humans rely on an understanding of the actions that can be performed on them, as well as their properties. When considering concrete motor actions, this knowledge has been called the object affordance. Can this notion be generalized to any type of interaction that one can have with an object? In this paper we introduce a method to represent objects in a space where each dimension corresponds to a broad mode of interaction, based on verb selectional preferences in text corpora. This object embedding makes it possible to predict human judgments of verb applicability to objects better than a variety of alternative approaches. Furthermore, we show that the dimensions in this space can be used to predict categorical and functional dimensions in a state-of-the-art mental representation of objects, derived solely from human judgements of object similarity. These results suggest that interaction knowledge accounts for a large part of mental representations of objects.

Keywords: affordance; object representation; embedding

Introduction

In order to interact with objects in our environment, we rely on an understanding of the actions that can be performed on them, and their dependence (or effect) on properties of the object. Gibson (1979) coined the term "affordance" to describe what the environment "provides or furnishes the animal". Norman (2013) developed the term to focus on the properties of objects that determine action possibilities. The notion of "affordance" emerges from the relationship between the properties of objects and human actions. If we consider "object" as meaning anything concrete that one might interact with in the environment, there will be thousands of possibilities, both animate and inanimate (see WordNet (Miller, 1998)). The same is true if we consider "action" as meaning any verb that might be applied to the noun naming an object (see VerbNet (Schuler, 2005)). Intuitively, only a relatively small fraction of all possible combinations of object and action will be plausible. Of those, many will also be trivial, e.g. "see" or "have" may apply to almost every object. Finally, different actions might reflect a similar mode of interaction, depending on the type of object they are applied to (e.g. "chop" and "slice" are distinct actions, but they are both used in food preparation).

Mental representations of objects encompass many aspects beyond function. Several studies (McRae, Cree, Seidenberg, & McNorgan, 2005; Devereux, Tyler, Geertzen, & Randall, 2014; Hovhannisyan et al., 2020) have normed thousands of binary properties for hundreds of objects. Properties could be taxonomic (category), functional (purpose), encyclopedic (attributes), or visual-perceptual (appearance), among others. While some properties were affordances in themselves, most reflected many affordances at once (e.g. "is a vegetable" means that it could be planted, cooked, sliced, etc).

Recently, Zheng, Pereira, Baker, and Hebart (2019) and M. Hebart, Zheng, Pereira, and Baker (2020) introduced SPoSE, a model of the mental representations of 1,854 objects in a 49-dimensional space. The model was derived from a dataset of 1.5M Amazon Mechanical Turk (AMT) judgments of object similarity, where subjects were asked which of a random triplet of objects was the odd one out. The model embedded each object as a vector in a space where each dimension was constrained to be sparse and positive. Triplet judgments were predicted as a function of the similarity between embedding vectors of the three objects considered. The authors showed that these dimensions were predictable as a *combination* of elementary properties in the (Devereux et al., 2014) norm that often co-occur across many objects. M. Hebart et al. (2020) further showed that 1) human subjects could coherently label what the dimensions were "about", ranging from categorical (e.g. is animate, food, drink, building) to functional (e.g. container, tool) or structural (e.g. made of metal or wood, has inner structure). Subjects could also predict what dimension values new objects would have, based on knowing the dimension value for a few other objects.

Our first goal is to produce an analogous "affordance embedding" for objects, where each dimension of the space groups together actions often applied to objects scoring high on that dimension. Our approach is based on the hypothesis that, if a set of verbs apply to the same objects, they apply for similar reasons. We compile applications of action verbs to nouns naming objects in large text corpora, and use the resulting dataset to produce an embedding. This embedding represents each object as a vector in a low-dimensional space, where each dimension groups verbs that apply to similar objects. Our second goal is to understand the degree to which affordance knowledge underlies the mental representation of objects, as instantiated in SPoSE. We do this by showing that most dimensions of the SPoSE representation of an object can be predicted from its affordance embedding, in particular those that are categorical or functional.

Related Work

The problem of determining, given an action and an object, whether the action can apply to the object was defined as *affordance mining* (Chao, Wang, Mihalcea, & Deng, 2015). The authors proposed complementary methods for solving the affordance mining problem by predicting a plausibility score for each combination of object and action. Subsequent work (Rubinstein, Levi, Schwartz, & Rappoport, 2015; Lucy & Gauthier, 2017; Forbes, Holtzman, & Choi, 2019; Utsumi, 2020) predicted properties of objects in the norms above from word embeddings (Mikolov, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014), albeit without a focus on affordances. In addition to object/action plausibility prediction, Ji, Shi, Guo, and Chen (2020) addressed the problem of determining whether a object1/action/object2 relationship was plausible. Other papers have focused on understanding the relevant visual features in objects that predict affordances (Myers, Teo, Fermüller, & Aloimonos, 2015; Sawatzky, Srikantha, & Gall, 2017; Wang & Tarr, 2020). This has been combined with text in robotics literature, but usually focusing on a restricted set of objects and manipulation actions. For validation of the rankings of verb applicability predicted by our model, we will use the datasets from (Chao et al., 2015) and (Wang & Tarr, 2020), as they are the largest available human rated datasets. In computational linguistics, P. S. Resnik (1993) introduced computational approaches to determining *selectional preference*, the degree to which a particular semantic class tends to be used as an argument to a given predicate. Several methods have been proposed to do this, leveraging some grouping of verbs and objects into classes (e.g. WordNet in (P. Resnik, 1996), or co-occurrence statistics of words in a corpus (Erk, 2007; Padó, Padó, & Erk, 2007; Séaghdha, 2010; VanDeCruys, 2014; Zhang et al., 2020). All of these methods could be used to score verbs by how applicable they are to a given noun, the ancillary task we use to make sure our embedding carries the relevant information. Our proposed embedding space is a latent variable model for verb-noun applications. While this is also the case for these papers, they would require extensive modification to add sparsity assumptions – important for interpretability – and to produce verb rankings.

Methods

Objects and Actions considered

We used the list of 1854 object concepts introduced in (M. N. Hebart et al., 2019) and for which SPoSE embeddings are available. This list sampled from concrete, picturable, and nameable nouns in American English, and was further expanded by an AMT study to elicit category[1] memberships. As we are not doing sense disambiguation for each noun that names an object, we will use "noun" or "object" interchangeably. We created our own verb list by having three annotators go through all verb categories on VerbNet (Schuler, 2005), and selecting those that included verbs that corresponded to an action[2] performed by a human on an object. We kept only those categories where all annotators agreed, and all verbs in each category. The resulting list has 2541 verbs.

Extraction of Verb Applications to Nouns from Text

We used the UKWaC and Wackypedia corpora (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008), with approximately, 2B and 1B tokens, and 88M and 43M sentences, respectively. The former is the result of a crawl of British web pages, while the latter is a subset of Wikipedia. Both have been cleaned and have clearly demarcated sentences, which is ideal for dependency parsing. We replaced all common bigrams in (Brysbaert, Warriner, & Kuperman, 2014) by a single token.

We identified all sentences containing both verbs and nouns in our list, and we used Stanza to produce dependency parses for them. We extracted all the noun-verb pairs in which the verb was a syntactic head of a noun having `obj` (object) or `nsubj:pass` (passive nominal subject) dependency relations. We compiled raw counts of how often each verb was used on each noun within a sentence, producing a count matrix M. Note that this is different from normal co-occurrence counts; those would register a count whenever verb and noun were both present within a short window (e.g. up to 5 words away from each other), *regardless* of whether the verb applied to the noun, or they were simply in the same sentence. Note also that the counts pertain to every possible meaning of the noun.

Finally, we converted the matrix M into a Positive Pointwise Mutual Information (PPMI (Turney & Pantel, 2010)) matrix P where, for each object i and verb k:

$$P(i,k) := \max\left(\log \frac{\mathbb{P}(M_{ik})}{\mathbb{P}(M_{i*}) \cdot \mathbb{P}(M_{*k})}, 0\right), \qquad (1)$$

$\mathbb{P}(M_{i*})$ and $\mathbb{P}(M_{*k})$ are marginal probabilities of i and k.

Object embedding in a verb usage space

Object embedding via matrix factorization Our embedding is based on a factorization of the PPMI matrix P (m objects by n verbs) into the product of matrices O (m objects by d dimensions) and V (n verbs by d dimensions), yielding $\widetilde{P} := OV^T \approx P$. O is the object embedding in d-dimensional space, and V is the weighting of each verb in each dimension. Each column $V_{:,k}$ of matrix V contains a pattern of verb usage for *dimension* k, which captures verb co-occurrence across all objects. Intuitively, if two verbs occur often with the same objects, they will both have high loadings on one of the d-dimensions; conversely, the objects they occur with will share high loadings on that dimension. The top-5 verb patterns for most of the 70 dimensions are shown in Table 1.

The idea of factoring a count matrix (or a transformation of it) dates back to Latent Semantic Analysis (Landauer &

[1] Main categories: food, animal, clothing, tool, drink, vehicle, fruit, vegetable, body part, toy, container, bird, furniture, sports equipment, musical instrument, dessert, part of car, weapon, plant, insect, kitchen tool, office supply, clothing accessory, kitchen appliance, home decor, medical equipment, and electronic device.

[2] Those VerbNet categories contained $\sim 10-50$ verbs sharing thematic roles and selectional preferences (e.g. fill-9.8, amalgamate-22.2, manner-speaking-37.3, build-26.1, remove-10.1, cooking-45.3, create-26.4, destroy-44, mix-22.1, vehicle-51.4.1, dress-41.1.1).

Table 1: Top 5 verbs in selected affordance dimensions.

Dimension	Top 5 verbs in each affordance dimension
1	invent, introduce, manufacture, develop
2	blanch, boil, steam, drain, cook
3	spot, observe, sight, hunt, watch
4	park, drive, hire, crash, rent
5	wield, grab, carry, hold, hand
6	squirt, formulate, dilute, smear, dissolve
7	capsize, moor, sail, beach, raft
8	grass, uproot, mulch, smother, clothe
9	wear, don, unbutton, match, button
10	coil, splice, braid, sever, thread
11	rouge, twinkle, flinch, twitch, sneer
12	mewl, breast, coo, breastfeed, swaddle
13	empty, fill, clean, clutter, line
14	tiptoe, totter, leer, yowl, mosey
15	serve, eat, cook, prepare, order
16	drink, sip, sup, swig, quaff
17	determine, compute, plot, ascertain
18	pasture, herd, slaughter, milk, tether
19	moo, pomade, gel, tweeze, primp
20	weave, drape, embroider, knit, sew
21	lob, hurl, fire, throw, explode
22	wet, moisten, rinse, soak, reuse
23	fleck, scallop, strew, emanate, pluck
24	sound, hear, play, blare, amplify
25	bare, swathe, waver, thump, tattoo
26	steal, recover, retrieve, discover, hide
27	freckle, moisturize, spritz, dehair, deflesh
28	close, open, shut, padlock, unlatch
29	sprinkle, mix, add, stir, blend
31	manufacture, buy, purchase, sell, design
32	dodder, skedaddle, snicker, roust, sober
33	extinguish, light, kindle, rekindle, flare
34	strangulate, fumble, glove, punt, bunt
35	unscrew, screw, slacken, disengage, tighten
36	declaw, leash, worm, feud, groom
37	hunt, kill, cull, exterminate, chase
38	unfasten, tighten, fasten, undo, loosen
39	dodder, skedaddle, snicker, roust, sober
40	deice, whir, flit, swagger, quiver
43	cloister, remarry, ostracize, unionize, intermarry
45	gabble, cluck, bridle, loll, lisp
47	winnow, mill, parboil, grind, reap
49	grill, baste, barbecue, marinate, brown
50	sharpen, blunt, wield, plunge, thrust
51	thicken, spoon, reheat, stir, simmer
52	sprain, hyperextend, flex, fracture, injure
54	eradicate, deter, swat, combat, discourage
57	cultivate, grow, plant, prune, propagate
58	pilot, board, rearm, crew, station
61	install, connect, disconnect, activate, operate
62	erect, carve, flank, adorn, construct
63	fish, catch, destress, whiff, degut
64	bake, leaven, ice, eat, serve
65	block, clog, dam, choke, flood
66	fit, mount, position, incorporate, attach
67	slice, peel, chop, dice, grate
68	unload, wheel, lug, load, transport
70	munch, scoff, eat, gobble, nibble

Dumais, 1997), and was investigated by many others (Turney & Pantel, 2010). If factorized into a product of two low-rank matrices, the structure of the matrix can be approximated while excluding noise or rare events. Given that the PPMI matrix P is positive, the matrices O and V are as well. We obtain them through a non-negative matrix factorization (NMF) problem

$$O^*, V^* = \underset{O,V}{\operatorname{argmin}} \|P - OV^T\|_F^2 + \beta \mathcal{R}(O,V), \qquad (2)$$

which can be solved through an iterative minimization procedure. For the regularization $\mathcal{R}(O,V)$, we chose the sparsity control $\mathcal{R}(O,V) \equiv \sum_{ij} O_{ij} + \sum_{ij} V_{ij}$. We used the NNDSVD initialization, a SVD-based initialization which favours sparsity on O and V and approximation error reduction. We found that the optimal dimensionality and sparsity were $d = 70$ and $\beta = 0.3$, respectively, using the two-dimensional hold-out cross validation procedure described in the Appendix. This procedure removes entire blocks of the matrix at a time, and reconstructs them using a decomposition of the rest of the matrix, using a range of dimensionality and sparsity settings.

Estimating the verb usage pattern for each object Deriving a similar pattern for each *object* i, given its embedding vector $O_{i,:} = [o_{i_1}, o_{i_2}, \ldots o_{i_d}]$, requires combining these patterns based on the weights given to each dimension. This requires computing the cosine similarity between each embedding dimension $O_{:,h}$ and the PPMI values $\tilde{P}_{:,k}$ for each verb k in the approximated PPMI matrix $\tilde{P} = OV^T$, which is

$$S(O_{:,h}, \tilde{P}_{:,k}) = \frac{O_{:,h} \cdot \tilde{P}_{:,k}}{\|O_{:,h}\|_2 \|\tilde{P}_{:,k}\|_2}. \qquad (3)$$

Given the embedding vector for object i, $O_{i,:} = [o_{i_1}, o_{i_2}, \ldots o_{i_d}]$, we compute the pattern of verb usage for the object as $O_{i,:}S$. Thus, this is a weighted sum of the similarity between every $O_{:,h}$ and $\tilde{P}_{:,k}$. We will refer to the ordering of verbs by this score as the *verb ranking* for object i.

Experiments and Results

Prediction of affordance plausibility

Affordance ranking task The first quantitative evaluation of our embedding focuses on the ranking of verbs as possible affordances for each object. We will use the Affordance Area Under The Curve (AAUC) relative to datasets that provide, for each object, a set of verbs known (or likely) to be affordances. Intuitively, the verb ranking for object i is good if it places these verbs close to the top of the ranking, yielding an AAUC of 1. Conversely, a random verb ranking would have an AAUC of 0.5, on average. This is a conservative measure, given that a perfect ranking would still penalize every true affordance not at the top. Hence, this is useful as a *relative* measure for comparing between our and competing approaches for producing rankings. More formally, given the K ground truth verb affordances $\{g_k\}_{k=1}^K$ of object i, and its verb ranking $\{v_i\}_{i=1}^n$, we denote ℓ_k to be the index such that $v_{\ell_k} = g_k \; \forall k$. We then define AUCC for object i as AUCC $= \frac{1}{K} \sum_{k=1}^K \left(1 - \frac{\ell_k}{n}\right)$.

Datasets We use the two largest publicly available object affordance datasets as ground truth. In the first dataset, WTAction (Wang & Tarr, 2020), objects are associated with the top 5 actions label provided by human annotators in response to "What can you do with this object?". Out of 1,046 objects and 445 actions in this dataset, there are 971 objects and 433 verbs that overlap with those in our lists ($\sim$ 3.12 action labels per object) . The second dataset, MSCOCO (Chao et al., 2015), scores every candidate action for an object ranging from 5.0 ("definitely an affordance") to 1.0 ("definitely not an affordance"). We consider only a 5.0 score as being an affordance. Out of 91 objects and 567 actions, 78 objects and 558 verbs overlap with ours ($\sim$ 34 action labels per object).

Baseline methods We compared the ranking of verbs produced by our algorithm with an alternative proposed in (Chao et al., 2015): ranking by the cosine similarity between word embedding vectors for each noun and those for all possible verbs in the dataset. We considered several off-the-shelf embedding alternatives, namely Word2Vec ((Mikolov et al., 2013), 6B token corpus), GloVe ((Pennington et al., 2014), 6B and 840B token corpora, Dependency-Based Word Embedding (DBWE, (Levy & Goldberg, 2014), 6B corpus), and Non-negative Sparse Embedding (NNSE, (Murphy, Pratim, & Tom, 2012), 16B corpus). The embeddings are 300-D in all cases, except for NNSE (1000-D, similar results for 2500-D). Finally, we also ranked the verbs by their values in the row of the PPMI matrix P for each probed object, to see how much our method of embedding through a low-rank approximation allowed the extraction of additional information.

Table 2: AAUC of verb rankings by each method.

Dataset	Method						
	DBWE	NNSE	W2V	GV	G840	LSA	Ours
WTA	0.60	0.65	0.70	0.75	0.80	0.81	0.88
MSC	0.56	0.58	0.59	0.65	0.68	0.63	0.77

Results For each dataset, we reduced our embeddings O and V according to the sets of objects and verbs available. We then obtained the verb ranking for each object, as described in the **Methods** section, as well as rankings predicted with the different baseline methods in the previous section. Table 2 shows average AAUC results obtained with these verb rankings on the two datasets. Our ranking is better than those of all the baseline methods, as well as PPMI (0.77, 0.61), as determined from paired two-sided t-tests, in both WTAction and MSCOCO (all p-values $\ll 0.01$). The following figure contrasts the AAUC distribution across objects for our method with those obtained with the top 4 embeddings and PPMI, on the WTaction and MSCOCO datasets, respectively.

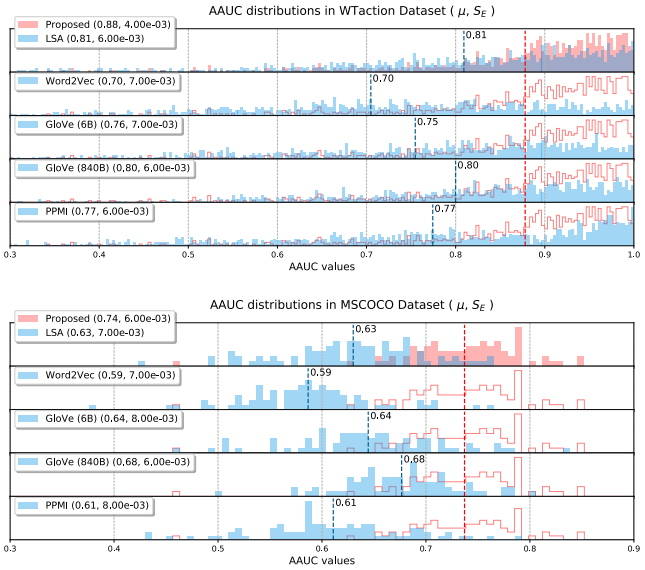

Figure 1: AAUC distribution on WTaction (top) and MSCOCO (bottom) datasets using our method, against the 4 top embeddings and PPMI.

Prediction of SPoSE object representations

The SPoSE representation and dataset The dimensions in the SPoSE representation (M. Hebart et al., 2020) are interpretable, in that human subjects coherently label what those dimensions are "about", from the categorical (e.g. animate, building) to the functional (e.g. can tie, can contain, flammable) or structural (e.g. made of metal or wood, has inner structure). The SPoSE vectors for objects are derived from behaviour in a "which of a random triplet of objects is the odd one out" task. The authors propose a hypothesis for why there is enough information in this data to allow this: when given any two objects to consider, subjects mentally sample the contexts where they might be found or used. The resulting dimensions reflect the aspects of the mental representation of an object that come up in that sampling process. The question we want to answer is, then, which of these dimensions reflect affordance or interaction information. We used the 49-D SPoSE embedding published with (M. Hebart et al., 2020). We excluded objects named by nouns that had no verb co-occurrences in our dataset and, conversely, verbs that had no interaction with any objects. We averaged the vectors for objects named by the same polysemous noun (e.g. "bat"). The resulting dataset had 1755 objects/nouns, and 2462 verbs.

Relationship between SPoSE and affordance dimensions We first considered the question of whether affordance dimensions correspond directly to SPoSE dimensions, by looking for the highest correlated match. Many of the 49 SPoSE dimensions are similar to at least one of the 70 affordance dimensions, with the distribution of correlation of the best match shown in the x-axis of Figure 2. Then, in order to determine which SPoSE dimensions of an object could be explained in terms of affordance dimensions, we predicted their value from the affordance embedding of the object. Denoting the SPoSE vectors for m objects as a $m \times 49$ matrix Y, we solved this Lasso regression problem for each column $Y_{:,i}$

$$w_i^* = \underset{w \in \mathbb{R}^d, w \geq 0}{\operatorname{argmin}} \frac{1}{2m} \|Y_{:,i} - Ow\|_2^2 + \lambda \|w\|_1, \quad i = 1, \ldots 49, \quad (4)$$

where λ was chosen based on a 2-Fold cross-validation, with λ in $[10e^{-7}, 10e^{3}]$ with log-scale spacing. Since both $Y_{:,i}$ and our embedding O represent object features by positive values, we restricted $w \geq 0$. Intuitively, this means that we try to explain every SPoSE dimension by combination of the *presence* of certain affordance dimensions, not by trading them off.

Overall, the cross-validated predictions of this regression model are more similar to SPoSE dimensions than any individual affordance dimension, as shown in the y-axis of Figure 2. The best predicted dimensions are categorical, e.g. "animal", "plant", or "tool", or functional, e.g. "can tie" or

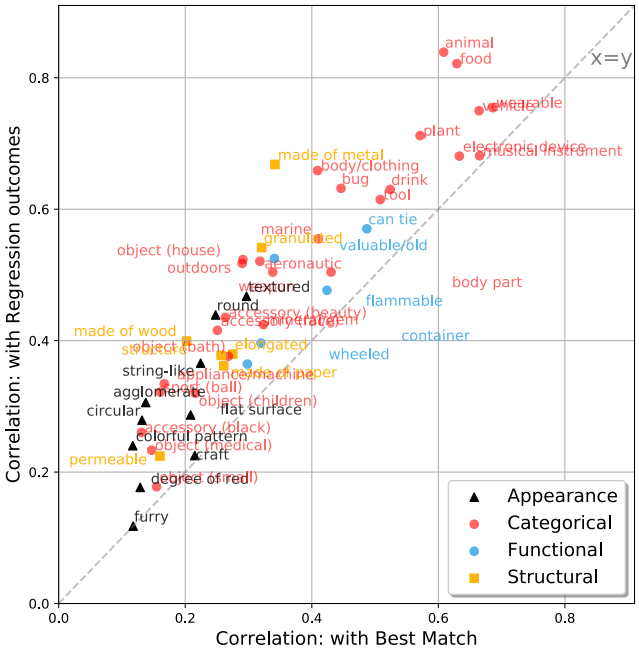

Figure 2: For each SPoSE dimension, correlation with the best matching affordance dimension (x-axis) and with the cross-validated prediction of the regression model for it (y-axis).

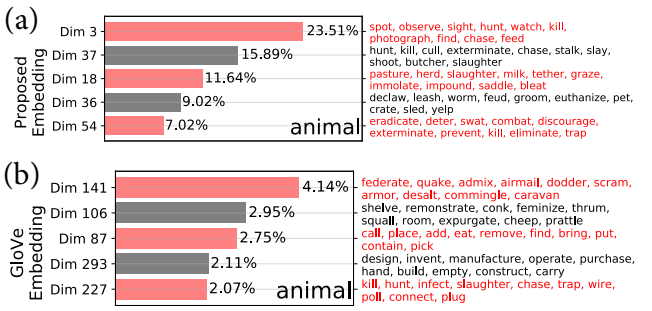

Figure 3: Top 10 verbs in the 5 most important affordance dimensions (proposed affordance embedding versus GloVe 840B) for predicting the "animal" SPoSE feature.

"flammable". Structural dimensions are also predictable, e.g. "made of metal", "made of wood", or "paper", but less so for appearance-related dimensions, e.g. "colorful pattern", "craft", or "degree of red". What can explain this pattern of predictability? Most SPoSE dimensions can be expressed as a linear combination of affordance dimensions, where both the dimensions and regression weights are *non-negative*. This leads to a sparse regression model – since dependent variables cannot be subtracted to improve the fit – where, on average, 5 affordance dimensions have 80% of the regression weight. Each affordance dimension, in turn, corresponds to a ranking over verbs. Figure 3a shows the top 10 verbs in the 5 most important affordance dimensions for predicting the "animal" SPoSE dimension. As each affordance dimension loads on verbs that correspond to broad modes of interaction (e.g. observation, killing, husbandry), the model is both predictive and interpretable. Whereas we could use dense embeddings to predict SPoSE dimensions, they do not work as well (in either accuracy or interpretability, see Figure 3b for GloVe). For example, if we consider the top 5 verbs from affordance dimensions used in predicting each SPoSE dimension, we see that "tool" has "sharpen, blunt, wield, plunge, thrust" (D50); "food" has "serve, eat, cook, prepare, order" (D15), or "bake, leaven, ice, eat, serve" (D64); "plant" shares D2 with "food", but also has "cultivate, grow, plant, prune, propagate" (D57).

These results suggest that SPoSE dimensions are predictable *insofar* as they can be expressed as combinations of modes of interaction with objects. As described in **Methods** section, we can combine affordance dimension verb rankings into a verb ranking for each SPoSE dimension. We replaced the embedding O in (3) with the SPoSE prediction $\widetilde{Y}$ and we ranked the verbs for dimension h according to $S(\tilde{Y}_{:,h}, \tilde{P}_k)$. Table 3 shows, for every SPoSE dimension, ranked by predictability, the top 10 verbs in its ranking. This table suggests that highly predictable categorical dimensions correspond to very clean affordances. The same is true for functional dimensions, e.g. "can tie" or "container" or "flammable"; even though they are not "classic" categories, subjects group items belonging to them based on their being suitable for a purpose (e.g. "fasten", "fill", or "burn"). Why would this hold for structural dimensions? One possibility is if objects having that dimension overlap substantially with a known category (e.g. "made of metal" and "tool"). Another is that the structure drives manual or mechanical affordance (e.g. "elongated" or "granulated"). Finally, what are the affordances for appearance dimensions that can be predicted? Primarily, actions on items in categories that share that appearance, e.g. "textured" is shared by fabric items, "round" is shared by many fruits or vegetables. Prediction is worse when the items sharing the dimension come from many different semantic categories.

Conclusions

In this paper, we introduced an approach to embed objects in a space where every dimension corresponds to a pattern of verb applicability to those objects. We view such a pattern as a very broad extension of the classical notion of "affordance", obtained by considering verbs that go well beyond concrete motor actions, and objects that encompass many different categories beyond tools or household objects. We showed that this embedding can be learned from a text corpus and used to rank verbs by how applicable they would be to a given object. We used our embedding to predict SPoSE dimensions for objects. This allowed us to conclude that our "affordance" embedding knowledge predicts 1) category information, 2) purpose, and 3) some structural aspects of the object. SPoSE dimensions to do with visual appearance were poorly predicted. To go beyond this, and conclude that our embedding is a valid model for mental representations of objects – insofar as our interactions with them go – would require additional experiments. One

Table 3: Affordance assignment for a selection of SPoSE dimensions mentioned in the text, ordered by how well they can be predicted from the affordance embedding. The names of SPoSE dimensions are simplified.

Correlation	SPoSE dimension	Type	Affordances (Top Ten Ranked Verbs)
0.84	animal	categorical	kill, spot, hunt, observe, chase, feed, slaughter, sight, trap, find
0.82	food	categorical	serve, eat, cook, prepare, taste, consume, add, mix, stir, order
0.75	wearable	categorical	wear, don, match, knit, sew, fasten, rip, embroider, tear, model
0.71	plant	categorical	grow, cultivate, plant, add, eat, chop, gather, cut, dry, prune
0.67	made of metal	structural	fit, invent, manufacture, incorporate, design, position, attach, utilize, carry, install
0.61	tool	categorical	wield, grab, hold, carry, sharpen, swing, hand, pick, clutch, throw
0.57	can tie	functional	fasten, tighten, unfasten, undo, attach, thread, tie, secure, loosen, loose
0.54	granulated	structural	contain, mix, scatter, add, gather, remove, sprinkle, dry, deposit, shovel
0.48	flammable	functional	light, extinguish, ignite, throw, carry, flash, kindle, place, manufacture, douse
0.47	textured	appearance	remove, place, hang, tear, stain, spread, weave, clean, drape, wrap
0.44	round	appearance	grow, cultivate, pick, add, slice, place, eat, chop, throw, plant
0.40	made of wood	structural	place, remove, carry, incorporate, design, contain, bring, construct, manufacture, find
0.40	container	functional	empty, fill, carry, place, clean, load, bring, dump, unload, leave
0.38	elongated	structural	grab, carry, wield, hold, pick, place, throw, hand, bring, drop
0.24	colorful pattern	appearance	manufacture, buy, design, place, remove, sell, invent, purchase, contain, bring
0.23	craft	appearance	place, bring, remove, design, hang, call, buy, put, pull, manufacture
0.22	permeable	structural	fit, incorporate, remove, place, design, manufacture, install, position, clean, attach
0.18	degree of red	appearance	place, call, add, contain, remove, find, buy, bring, introduce, sell

possibility would be to explicitly ask human subjects "given objects that load highly on this embedding dimension, what can you do with them", and consider the typicality of verb answers against the weight given to those verbs by the dimension. Given that our embedding is based on language data about which verbs apply to which objects, we would expect these experiments to give verb loadings coherent with ours.

A future direction of work will be to predict SPoSE dimensions that are not well explained in terms of affordance embeddings. We plan to do this using embeddings produced with the same framework, but from different co-occurrence statistics. The first possibility will be to extract instances in corpora where objects are the *subjects* of verbs, i.e. they act or cause certain effects. The second possibility will be to consider applications of adjectives to objects, given that those may contain information relevant to all 4 types of SPoSE dimensions. Finally, we will consider reducing visual representations of objects obtained through deep neural networks to embedding vectors, as those contain both visual and semantic information.

Acknowledgments This work was supported by the National Institute of Mental Health Intramural Research Program (ZIC-MH002968, ZIA-MH002035). This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The authors would like to thank Martin Hebart and Charles Zheng for patiently sharing SPoSE and THINGS resources with us, and Aria Wang for graciously giving us access to her object affordance dataset.

References

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Chao, Y.-W., Wang, Z., Mihalcea, R., & Deng, J. (2015). Mining semantic affordances of visual object categories. In *Proceedings of the IEEE CVPR* (pp. 4259–4267).

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (CSLB) concept property norms. *BRM*, *46*(4), 1119–1127.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. *ACL 2007*(June), 216–223.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of english. In *Wac-4* (pp. 47–54).

Forbes, M., Holtzman, A., & Choi, Y. (2019). Do neural language representations learn physical commonsense? *arXiv:1908.02899*.

Gibson, J. J. (1979). *Ecological approach to visual percept.*

Hebart, M., Zheng, C. Y., Pereira, F., & Baker, C. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgments.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10).

Hovhannisyan, M., Geib, B., Clarke, A., Cicchinelli, R., Cabeza, R., & Davis, S. (2020). The visual and semantic features that predict object memory: Concept property norms for 1000 object images.

Ji, L., Shi, B., Guo, X., & Chen, X. (2020). Functionality discovery and prediction of physical objects. In *AAAI 2020* (Vol. 34, pp. 123–130).

Kanagal, B., & Sindhwani, V. (2010). Rank selection in low-rank matrix approximations: A study of cross-validation for NMFs. In *NIPS 2010* (Vol. 1, pp. 10–15).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *NIPS 2001* (pp. 556–562).

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *ACL 2014* (pp. 302–308).

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *arXiv:1705.11168*.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set

of living and nonliving things. *BRM*, *37*(4), 547–559.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.
Miller, G. A. (1998). *Wordnet: An electronic lexical database*. MIT press.
Murphy, B., Pratim, P., & Tom, T. (2012). Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *ACL 2021*.
Myers, A., Teo, C. L., Fermüller, C., & Aloimonos, Y. (2015). Affordance detection of tool parts from geometric features. In *ICRA 2015* (pp. 1374–1381).
Norman, D. (2013). *Design of everyday things*. Basic books.
Padó, S., Padó, U., & Erk, K. (2007). Flexible, corpus-based modelling of human plausibility judgements. *EMNLP-CoNLL 2007*(June), 400–409.
Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP 2014* (pp. 1532–1543).
Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*(1-2), 127–159.
Resnik, P. S. (1993). Selection and information: a class-based approach to lexical relationships. *IRCS TR Series*, 200.
Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge? In *ACL 2015* (pp. 726–730).
Sawatzky, J., Srikantha, A., & Gall, J. (2017). Weakly supervised affordance detection. In *IEEE CVPR 2017* (pp. 2795–2804).
Schuler, K. K. (2005). Verbnet: A broad-coverage, comprehensive verb lexicon.
Séaghdha, D. Ó. (2010). Latent variable models of selectional preference. *ACL 2010*(July), 435–444.
Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *JAIR*, *37*, 141–188.
Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, *44*(6), e12844.
VanDeCruys, T. (2014). A neural network approach to selectional preference acquisition. *EMNLP 2014*, 26–35.
Wang, A. Y., & Tarr, M. J. (2020). Learning intermediate features of object affordances with a convolutional neural network. *arXiv:2002.08975*.
Zhang, H., Bai, J., Song, Y., Xu, K., Yu, C., Song, Y., … Yu, D. (2020). Multiplex word embeddings for selectional preference acquisition. *arXiv:2001.02836*.
Zheng, C. Y., Pereira, F., Baker, C., & Hebart, M. (2019). Revealing interpretable object representations from human behavior. In *ICLR 2019*. OpenReview.net.

Appendix: Hyper-parameter Selection for Non-negative Matrix Factorization

Denote $M_t, M_v \in \{0,1\}^{n\times m}$ to be the mask matrices for indicating held-in and held-out entries of the input PPMI matrix P

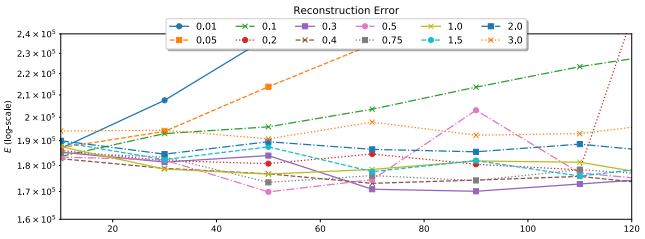

Figure 4: A zoom-in plot for the reconstruction errors.

in CV procedure, we then optimize for O^* and V^*:

$$O^*, V^* = \underset{O,V}{\operatorname{argmin}} \|M_t \odot (P - OV^T)\|_F^2 + \beta \mathcal{R}(O,V). \tag{5}$$

To apply the multiplicative update scheme in (Lee & Seung, 2001), we need the partial derivatives with respect to O and V. Denote $F(O,V) \equiv \|M_t \odot (P - OV^T)\|_F^2 + \beta \mathcal{R}(O,V)$, we have

$$\begin{aligned} \nabla_O F(O,V) &= (M_t \odot OV^T)V - (M \odot P)V + \beta \cdot \mathbf{1} \\ \nabla_V F(O,V) &= (M_t \odot OV^T)^T U - (M \odot P)^T U + \beta \cdot \mathbf{1}. \end{aligned} \tag{6}$$

We then have the following update rules that is guaranteed to be non-increasing:

$$\begin{aligned} O^{(i+1)} &\leftarrow O^{(i)} \odot \frac{(M_t \odot P)V^{(i)}}{(M_t \odot O^{(i)}(V^{(i)})^T)V^{(i)} + \beta} \\ V^{(i+1)} &\leftarrow V^{(i)} \odot \frac{(M_t \odot P)^T U^{(i)}}{(M_t \odot O^{(i)}(V^{(i)})^T)^T U^{(i)} + \beta}, \end{aligned} \tag{7}$$

where the fraction here represents elementary-wise division. For the choice of M_t and M_v, we follow the same approach as proposed in (Kanagal & Sindhwani, 2010). We first split the matrix into K blocks, with randomly shuffled rows and columns. Denote $\mathtt{r}^{(k)}$ and $\mathtt{c}^{(k)}$ to be the index vectors for rows and columns respectively, where $\mathtt{r}_i^{(k)} = 1$ if row i is in block k, or $\mathtt{c}_j^{(k)} = 1$ if column j is in block k. The mask for k-th block can then be expressed as $M^{(k)} = \mathtt{r}^{(k)} \otimes \mathtt{c}^{(k)}$. We then randomly select q out of K blocks as holdout blocks, which gives

$$M_v = \sum_{s=1}^{q} \mathtt{r}^{(k_s)} \otimes \mathtt{c}^{(k_s)}, \quad M_t = \mathbf{1} - M_v, \tag{8}$$

where k_s is the index of selected block. The reconstruction error E can thus be computed:

$$E = \|M_v \odot (P - O^*(V^*)^T)\|_F^2 + \beta \mathcal{R}(O^*, V^*). \tag{9}$$

Figure 4 shows a zoom-in plot of the reconstruction error under different combinations of d and β. For every (d, β) setting, we perform multiple optimization since NMF is sensitive to initialization. We then choose $d = 70$ and $\beta = 0.3$ accordingly. Empirically, we observe that the rank selection is quite robust to over-fitting when there is a sufficient sparsity control, for instance, $\beta > 0.1$ in our dataset. We also observe that whenever $d \in [50, 150]$ and $\beta \in [0.05, 0.5]$, the results are similar.