UNDERSTANDING MENTAL REPRESENTATIONS OF OBJECTS THROUGH VERBS APPLIED TO THEM

Anonymous authors
Paper under double-blind review

ABSTRACT

In order to interact with objects in our environment, we rely on an understanding of the actions that can be performed on them, and the extent to which they rely or have an effect on the properties of the object. This knowledge is called the object "affordance". We propose an approach for creating an embedding of objects in an affordance space, in which each dimension corresponds to an aspect of meaning shared by many actions, using text corpora. This embedding makes it possible to predict which verbs will be applicable to a given object, as captured in human judgments of affordance, better than a variety of alternative approaches. Furthermore, we show that the dimensions learned are interpretable, and that they correspond to typical patterns of interaction with objects. Finally, we show that the dimensions can be used to predict a state-of-the-art mental representation of objects, derived purely from human judgements of object similarity.

1 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 (2014) 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 the 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 et al., 2005; Devereux et al., 2014; Hovhannisyan et al., 2020) have asked human subjects to list binary properties for hundreds of objects, yielding thousands of answers. Properties could be taxonomic (category), functional (purpose), encyclopedic (attributes), or visual-perceptual (appearance), among other groups. While some properties were affordances in themselves (e.g. "edible"), others reflected many affordances at once (e.g. "is a vegetable" means that it could be planted, cooked, sliced, etc). More recently, Zheng et al. (2019); Hebart et al. (2020) introduced SPoSE, a model of the mental representations of objects. 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 was an embedding for objects where each dimension was constrained to be sparse and positive, and where 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. 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. SPoSE is unusual in its wide coverage – 1,854 objects – and in having been validated in independent behavioral data.

Our first goal is to produce an analogous affordance embedding space for objects, where each dimension groups together actions corresponding to a particular ”mode of interaction”. Our second goal is to understand the degree to which affordance knowledge underlies the mental representation of objects, as instantiated in SPoSE. In this paper, we will introduce and evaluate an approach for achieving both of these goals. Our approach is based on the hypothesis that, if a set of verbs apply to the same objects, they apply for similar reasons. We start by identifying applications of action verbs to nouns naming objects, in large text corpora. We then use the resulting dataset to produce an embedding that represents each object as a vector in a low-dimensional space, where each dimension groups verbs that tend to be applied to similar objects. We do this for larger lists of objects and action verbs than previous studies (thousands in each case). Combining the weights on each verb assigned by various dimensions yields a ranking over verbs for each concept. We show that this allows us to predict which verbs will be applicable to a given object, as captured in human judgments of affordance. Further, we show that the dimensions learned are interpretable, and they group together verbs that would all typically occur during certain complex interactions with objects. Finally, we show that they can be used to predict most dimensions of the SPoSE representation, in particular those that are categorical or functional. This suggests that affordance knowledge underlies much of the mental representation of objects, in particular semantic categorization.

2 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” in Chao et al. (2015). The authors proposed complementary methods for solving the affordance mining problem by predicting a plausibility score for each combination of object and action. The best method used word co-occurrences in two ways: n-gram counts of verb-noun pairs, or similarity between verb and noun vectors in Latent Semantic Analysis (Deerwester et al., 1990) or Word2Vec (Mikolov et al., 2013) word embeddings. For evaluation, they collected AMT judgements of plausibility (“is it possible to $< verb >$ a $< object >$”) for every combination of 91 objects and 957 action verbs. The authors found they could retrieve a small number of affordances for each item, but precision dropped quickly with a higher recall.

Subsequent work (Rubinstein et al., 2015; Lucy & Gauthier, 2017; Utsumi, 2020) predicted properties of objects in the norms above from word embeddings (Mikolov et al., 2013; Pennington et al., 2014), albeit without a focus on affordances. Forbes et al. (2019) extracted 50 properties (some were affordances) from Devereux et al. (2014), for a set of 514 objects, to generate positive and negative examples for 25,700 combinations. They used this data to train a small neural network to predict these properties. The input to the network was either the product of the vectors for object and property, if using word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Levy & Goldberg, 2014), or the representation of a synthesized sentence combining them, if using contextualized embeddings (Peters et al., 2018; Devlin et al., 2018). They found that the latter outperformed the former for property prediction, but none allowed reliable affordance prediction. In addition to object/action plausibility prediction, Ji et al. (2020) addressed the problem of determining whether a object1/action/object2 (target of the action with object1) was feasible. They selected a set of 20 actions from Chao et al. (2015) and combined them with the 70 most frequent objects in ConceptNet (Speer & Havasi, 2012) into 1400 object/action pairs, which were then labelled as plausible or not; given rater disagreements, this yielded 330 positive pairs and 1070 negative ones. They then combined the positive pairs with other objects as ”tails” (recipients of the action), yielding 3900 triplets. They reached F1 scores of 0.81 and 0.52 on the two problems, respectively. Other papers focus on understanding the relevant visual features in objects that predict affordances, e.g. (Myers et al., 2015; Sawatzky et al., 2017; Wang & Tarr, 2020). The latter collected affordance judgments on AMT (“what can you do with $< object >$”) for 500 objects and harmonized them with WordNet synsets for 334 action verbs.

We are not aware of any related work on representing objects in a space where dimensions reflect verb applicability, or where predictions cover thousands of objects and 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.

3 DATA AND METHODS

3.1 OBJECTS AND ACTIONS CONSIDERED

In this paper, we use the list of **1854** object concepts introduced in (Hebart et al., 2020). This sampled systematically from concrete, picturable, and nameable nouns in American English, and was further expanded by a crowdsourced study to elicit category membership ratings. Note that the list includes many things besides small objects, e.g. animals, insects, vehicles. We will refer to objects and the nouns naming them interchangeably. We created a 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 performed by a human on an object. We kept all the verbs in each selected category, and selected only those categories where all annotators agreed. The resulting list has **2541** verbs.

3.2 EXTRACTION OF VERB APPLICATIONS TO NOUNS FROM TEXT CORPORA

We used the UKWaC and Wackypedia corpora (Ferraresi et al., 2008) containing, 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 extensively cleaned up and have clearly demarcated sentences, which makes them ideal for dependency parsing. We replaced all common bigrams in Brysbaert et al. (2014) by a single token. We identified all sentences containing both verbs and nouns in our list, and we used Stanza (Qi et al., 2020) 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, producing a count matrix M. Note that this is very 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. Out of the 1854 nouns considered, there were 1755 with at least one associated action, and this is the list we will use in the remainder of this paper. Note also that the counts pertain to every possible meaning of the noun, given that no word sense disambiguation was performed.

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:

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

where $\mathbb{P}(M_{i*})$ and $\mathbb{P}(M_{*k})$ are respectively the marginal probability of i and k. P can be viewed as a pair-pattern matrix (Lin & Pantel, 2001), where the PPMI helps in separating frequency from informativess of the co-occurrence of nouns and verbs (Turney & Pantel, 2010; Turney & Littman, 2003). However, PPMI is also biased and may be large for rare co-occurrences, e.g. for (o_i, v_k) that co-occur only once in M. This is addressed in the process described in the next section.

3.3 OBJECT EMBEDDING IN A VERB USAGE SPACE

Object embedding via matrix factorization Our object embedding is based on a factorization of the PPMI matrix P (m objects by n verbs) into the product of low-rank matrices O (m objects by d dimensions) and V (n verbs by d dimensions), yielding an approximation $\widetilde{P} := OV^T \approx P$. O is the object embedding in d-dimensional space, and V is the verb loading for each of the d dimensions. 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 idea of factoring a count matrix (or a transformation of it) dates back to Latent Semantic Analysis (Landauer & Dumais, 1997), and was investigated by many others ((Turney & Pantel, 2010) is a good review). Given that PPMI is biased towards rare pairs of noun/verb, the matrix P is not necessarily very sparse. If factorized into a product of two low-rank matrices, however, the structure of the matrix can be approximated while excluding noise or rare events (Bullinaria & Levy, 2012).

Optimization problem Given that PPMI is positive, the matrices O and V are as well. This allows us to obtain them through a non-negative matrix factorization (NMF) problem

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

which can be solved through an iterative minimization procedure. For the regularization $\mathcal{R}(O, V)$, we chose the sparsity control $\mathcal{R}(O, V) \equiv \sum_{ij} O_{ij} + \sum_{ij} V_{ij}$.

In our experiments, we use $d = 70$ and $\beta = 0.3$. These values were found using the two-dimensional hold-out cross validation (Kanagal & Sindhwani, 2010), due to its scalability and natural fit to the multiplicative update algorithm for solving (2). Specifically, denoting M_t and M_v to be the mask matrices representing the held-in and held-out entries, we optimize for

$$O^*, V^* = \underset{O,V}{\operatorname{argmin}} \|M_t \odot (P - OV^T)\|_F^2 + \beta\mathcal{R}(O, V) \tag{3}$$

and obtain the reconstruction error $E = \|M_v \odot (P - O^*(V^*)^T)\|_F^2 + \beta\mathcal{R}(O^*, V^*)$. For more details, please refer to Appendix A. The optimization problem (2) is NP-hard and all state-of-the-art algorithms may converge only to a local minimum (Gillis, 2014); choosing a proper initialization of O and V is crucial. We used the NNDSVD initialization (Boutsidis & Gallopoulos, 2008), a SVD-based initialization which favours sparsity on O and V and approximation error reduction.

Estimating the verb usage pattern for each object Each column $V_{:,k}$ of matrix V contains a pattern of verb usage for *dimension* k, which captures verb co-occurrence across all objects. 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. The first step in doing so 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

$$S(O_{:,h}, \tilde{P}_{:,k}) = \frac{O_{:,h} \cdot \tilde{P}_{:,k}}{\|O_{:,h}\|_2 \|\tilde{P}_{:,k}\|_2}. \tag{4}$$

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.

4 EXPERIMENTS AND RESULTS

4.1 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. Note that 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 $\text{AUCC} = \frac{1}{K} \sum_{k=1}^K \left(1 - \frac{\ell_k}{n}\right)$.

Datasets We used 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") marked by 5 different workers. We consider only a 5.0 score as being an affordance, whereas Chao et al. (2015) used both 4.0 and 5.0. Out of 91 objects and 567 actions, there are 78 objects and 558 verbs that 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, and our 2B corpus), and Dependency-Based Word Embedding (Levy & Goldberg (2014), 6B corpus). Finally, we included the other two methods in (Chao et al., 2015), LSA (Deerwester et al. (1990), trained on our corpus), and ranking by frequency of verb/noun pair in Google N-grams (Lin et al. (2012)). The embeddings are 300-D in all cases. We contrasted Word2Vec and GloVe because they are based on two different embedding approaches (negative sampling and decomposition of a word co-occurrence matrix), developed on corpora twice as large as ours. We contrasted 6B and 840B versions of GloVe to see the effect of increasing dataset size, and 2B to show the effect of our corpus. In these methods, the co-occurrence considered is simply proximity within a window of a few tokens, rather than application of the verb to the noun. We included Levy & Goldberg (2014) as it is the only embedding using dependency parse information (albeit to define the word-cooccurrence window, rather than select verb applications to nouns as we do). We also ranked the verbs by their values in the row of the PPMI matrix P corresponding to each probed object, to see the effect of using a low-rank approximation in extracting information.

Table 1: Average AAUC of verb rankings produced by our and baseline methods

Method Dataset		LSA	DBWE	N-gram	W2V (6B)	GloVe (2B)	GloVe (6B)	GloVe (840B)	PPMI	Ours
WTaction	μ	0.81	0.60	0.65	0.70	0.79	0.75	0.80	0.77	**0.88**
	S_E	5e−3	6e−3	8e−3	7e−3	7e−3	7e−3	6e−3	6e−3	**4e−3**
MSCOCO	μ	0.63	0.56	0.58	0.59	0.66	0.64	0.68	0.61	**0.77**
	S_E	7e−3	8e−3	8e−3	7e−3	7e−3	8e−3	6e−3	8e−3	**5e−3**

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 Section 3.3, as well as rankings predicted with the different baseline methods in the previous section. Table 1 shows the AAUC results (in terms of the mean μ and the standard error S_E of AAUCs) obtained with these verb rankings on the two datasets. Our ranking is better than those of all the baseline methods, as determined from a paired two-sided t-test, in both WTAction (p-values of 7.36e−23, 6.82e−174, 1.34e−53, 1.44e−89, 5.21e−47, 8.70e−25 and 8.14e−39) and MSCOCO (p-values of 1.54e−23, 9.32e−36, 2.15e−27, 2.82e−30, 9, 64e−20, 7.93e−17 and 2.13e−29) datasets.

The overall distributions of AAUCs for the two datasets are shown in Appendix B. Our procedure yields AAUC closer to 1.0 for many more items than the other methods. This suggests that the co-occurrence of verb and noun within a window of a few tokens, the basis of the word embeddings that we compare against, carries some information about affordances, but also includes other relationships and noise (e.g. if the verb is in one clause and the noun in another). Results are better in the embedding trained in a corpus 280X larger, but still statistically worse than those of our procedure. Ranking based on the PPMI matrix P performs at the level of the 6B token embeddings. This suggests that our procedure is effective at removing extraneous information in the matrix P.

4.2 PREDICTION OF SPOSE OBJECT REPRESENTATIONS FROM OUR AFFORDANCE EMBEDDING

The SPoSE representation and dataset The dimensions in the SPoSE representation (Hebart et al., 2020) are interpretable, in that human subjects coherently label what those dimensions are "about", ranging 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 solely 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 natural question is, then, which of these dimensions reflect affordance information, and that is what our experiments aim to answer.

For our experiments, we used the 49-D SPoSE embedding published with (Hebart et al., 2020). Out of these, 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 each set of objects named by the same polysemous noun (e.g. "bat"). This resulted in a dataset of 1755 objects/nouns, with their respective SPoSE embedding vectors, and 2462 verbs.

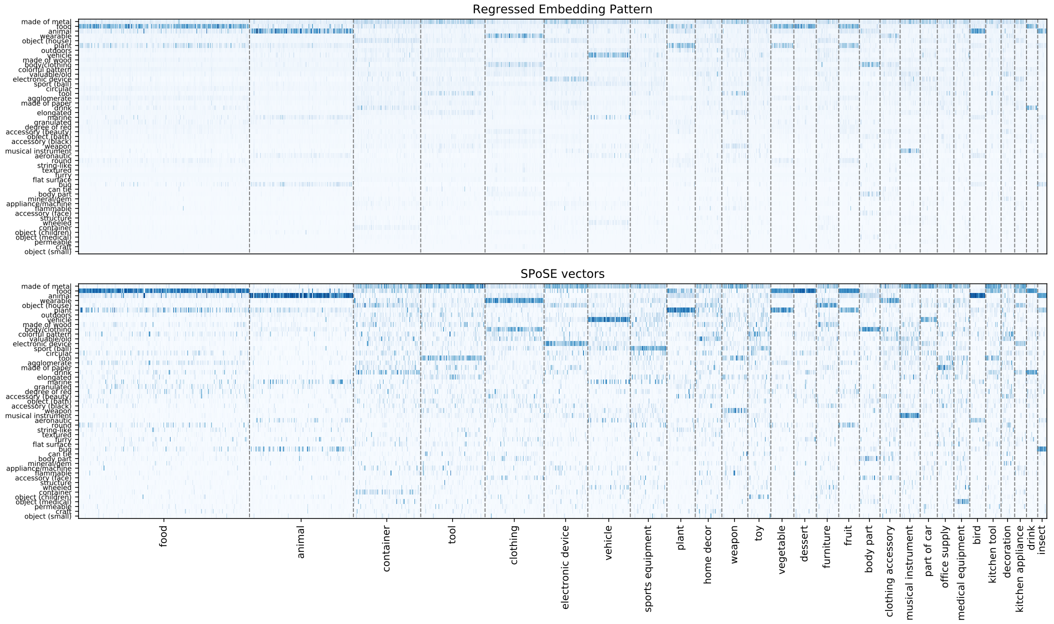


Figure 1: Prediction of SPoSE embeddings from affordance embeddings (top) versus actual SPoSE embeddings (bottom). Objects are grouped by semantic category (those with ≥ 15 objects).

Prediction of SPoSE meaning dimensions from affordance embedding The first experiment we carried out was to predict the SPoSE dimensions for an object from its representation in our affordance embedding. The ability to do this tells us which SPoSE dimensions – and, crucially, which *kinds* of SPoSE dimensions – can be explained in terms of affordances, to some extent.

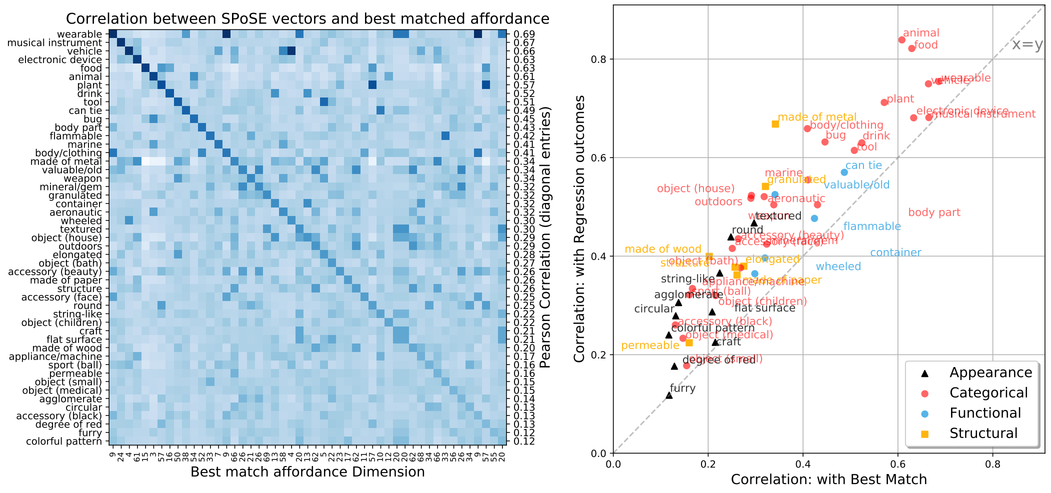
Denoting the SPoSE embeddings for m objects as a $m \times 49$ matrix Y, we solved the following Lasso regression problem for each column $Y_{:,i}$

$$w_i^* = \arg \min_{w \in \mathbb{R}^d, w \geq 0} \frac{1}{2m} \|Y_{:,i} - Ow\|_2^2 + \lambda \|w\|_1, \quad i = 1, 2, \ldots 49, \tag{5}$$

where λ was chosen based on a 2-Fold cross-validation, with λ in $[10\mathrm{e}^{-7}, 10\mathrm{e}^{3}]$ 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.

Figure 1 shows the predictions $\tilde{Y}_{:,i} = Ow_i^*$ and true values $Y_{:,i}$ for each SPoSE dimension i. For clarity, objects are grouped by their semantic category and only plotted for categories with size ≥ 15. The visual resemblance of the patterns, and the range of correlations between true and predicted dimensions (top: 0.84, mean: 0.46), see Figure 2b for distribution), indicate that the affordance embedding contains enough information to predict most SPoSE dimensions.

Relationship between SPoSE and affordance dimensions We first considered the question of whether affordance dimensions correspond directly to SPoSE dimensions, by looking for the best match for each of the latter in terms of correlation. As shown in Figure 2(a), most SPoSE dimensions have one or more affordance dimensions that are similar to them. However, when we consider the cross-validated regression models to predict SPoSE dimensions from affordance dimensions, *every* model places non-zero weight on several of the latter. Furthermore, those predictions are much more similar to SPoSE dimensions than almost any individual affordance dimension. Figure 2(b) plots, for every SPoSE dimension, the correlation with its closest match (x-axis) versus the correlation



(a) Correlation with outcomes

(b) Comparison

Figure 2: **a)** Correlation between each SPoSE dimension and the corresponding best match in our affordance embedding (correlation values shown in right vertical axis). **b)** For each SPoSE dimension, the relationship between the similarity with the best matching affordance dimension (x-axis) and the similarity with the cross-validated prediction of the regression model for it (y-axis).

with the cross-validated prediction of the regression model (y-axis). The best predicted dimensions are categorical, e.g. "animal", "plant", or "tool", or functional, e.g. "can tie" or "flammable". Structural dimensions are also predictable, e.g. "made of metal", "made of wood", or "paper", but appearance-related dimensions, e.g. "colorful pattern", "craft", or "degree of red", less so.

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 coherent modes of interaction (e.g. observation, killing, husbandry), the model is not only predictive but also interpretable. Whereas we could also use dense embeddings to predict SPoSE dimensions, they do not work as well (in either accuracy or interpretability, see Figure 3b for GloVe).

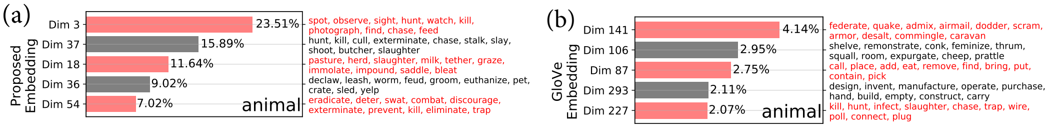


Figure 3: Top 10 verbs in the 5 most important affordance dimensions (proposed affordance embedding versus GloVe 840B) for predicting the "animal" SPoSE feature

If we consider the top 5 verbs from affordance dimensions that are most used in predicting each SPoSE dimension, we see that "tool" has "sharpen, blunt, wield, plunge, thrust" (D50); "animal" has "spot, observe, sight, hunt, watch" (D3), "hunt, kill, cull, exterminate, chase" (D37), or "pasture, herd, slaughter, milk, tether" (D18); "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). The full list of affordance dimensions most relevant for predicting each SPoSE dimension is provided in Appendix E (Figure 7–12).

These results suggest that SPoSE dimensions are predictable *insofar* as they can be expressed as combinations of modes of interaction with objects. Using the approach in Section 3.3, we produced a combined ranking over verbs for each SPoSE dimension. We first replaced the embedding O in (4) with the SPoSE prediction $\widetilde{Y}$ and we ranked the verbs for dimension h according to $S(\tilde{Y}_{:,h}, \tilde{P}_k)$. Table 2 shows, for every SPoSE dimension, ranked by predictability, the top 10 verbs in its ranking. Examining this table provides empirical confirmation that, first and foremost, 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"). But why would this hold for structural dimensions? One possible reason is that 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 appears to be worse when the items sharing the dimension come from many different semantic categories (judged with the THINGS (Hebart et al., 2019) database, which contains the pictures of items shown to subjects).

Pearson correlation	Dimension label	Taxonomy	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

Table 2: Affordance assignment for a selection of SPoSE dimensions mentioned in the text, ordered by how well they can be predicted from the affordance embedding. Dimension labels are simplified. The full table and descriptions for each dimension are provided in Appendices C and D, respectively.

5 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 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 evaluated this prediction against two separate human judgment ground truth datasets, and verified that our method outperforms general-purpose word embeddings trained on much larger text corpora. Given this validation of the embedding, we were able to use it to predict SPoSE dimensions for objects. This is an embedding that was derived from human behavioural judgements of object similarity and has interpretable dimensions, which have been validated in psychological experiments. We used the resulting prediction models to probe the relationship between affordances (as patterns of verb co-occurrence over many objects) and the various types of SPoSE dimensions. This allowed us to conclude that affordance knowledge predicts 1) category information, 2)purpose, and 3) some structural aspects of the object. To increase prediction quality in future work, one approach will be to enrich and refine the co-occurrence matrix in larger corpora, now that the basic approach has been shown to be feasible. Another interesting future direction will be to understand how affordance could be driven by more fine-grained visual appearance properties, by considering other semantic dependencies, or jointly using text and image features such as (Wang & Tarr, 2020).

REFERENCES

Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, 41(4):1350–1362, 2008.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911, 2014.

John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, 44(3):890–907, 2012.

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4259–4267, 2015.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41 (6):391–407, 1990.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (CSLB) concept property norms. *Behavior research methods*, 46(4): 1119–1127, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. Introducing and evaluating ukWaC, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pp. 47–54, 2008.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*, 2019.

James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

Nicolas Gillis. The why and how of nonnegative matrix factorization. *Regularization, optimization, kernels, and support vector machines*, 12(257):257–291, 2014.

Martin Hebart, Charles Y Zheng, Francisco Pereira, and Chris Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. 2020.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10), 2019.

Mariam Hovhannisyan, Benjamin Geib, Alex Clarke, Rosalie Cicchinelli, Roberto Cabeza, and Simon Davis. The visual and semantic features that predict object memory: Concept property norms for 1000 object images. 2020.

Lei Ji, Botian Shi, Xianglin Guo, and Xilin Chen. Functionality discovery and prediction of physical objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 123–130, 2020.

Bhargav Kanagal and Vikas Sindhwani. Rank selection in low-rank matrix approximations: A study of cross-validation for NMFs. In *Proc Conf Adv Neural Inf Process*, volume 1, pp. 10–15, 2010.

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

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 302–308, 2014.

Dekang Lin and Patrick Pantel. DIRT–Discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 323–328, 2001.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, and Slav Petrov. Syntactic annotations for the google books ngram corpus. 2012.

Li Lucy and Jon Gauthier. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *arXiv preprint arXiv:1705.11168*, 2017.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1374–1381. IEEE, 2015.

Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages, 2020.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 726–730, 2015.

Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2795–2804, 2017.

Karin Kipper Schuler. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005.

Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC*, pp. 3679–3686, 2012.

Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

Akira Utsumi. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844, 2020.

Aria Yuan Wang and Michael J Tarr. Learning intermediate features of object affordances with a convolutional neural network. *arXiv preprint arXiv:2002.08975*, 2020.

Charles Y. Zheng, Francisco Pereira, Chris I. Baker, and Martin N. Hebart. Revealing interpretable object representations from human behavior. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=ryxSrhC9KX`.