VICE: Variational Interpretable Concept Embeddings

Anonymous Author(s)
Affiliation
Address
email

Abstract

A central goal in the cognitive sciences is the development of numerical models for mental representations of object concepts. This paper introduces Variational Interpretable Concept Embeddings (VICE), an approximate Bayesian method for embedding object concepts in a vector space using data collected from humans in an odd-one-out triplet task. VICE uses variational inference to obtain sparse, non-negative representations of object concepts with uncertainty estimates for the embedding values. These estimates are used to automatically select the dimensions that best explain the data. We derive a PAC learning bound for VICE that can be used to estimate generalization performance or determine sufficient sample size in experimental design. VICE rivals or outperforms its predecessor, SPoSE, at predicting human behavior in the odd-one-out triplet task. Furthermore, VICE's object representations are more reproducible and consistent across random initializations.

1 Introduction

Human knowledge about object concepts encompasses many types of information, including function or purpose, visual appearance, encyclopedic facts, or taxonomic characteristics. A central question in cognitive science concerns the representation of this knowledge and its use across different tasks. One approach to this question is inductive and lets subjects list properties for hundreds to thousands of objects [27, 9, 6, 19], yielding large lists of responses about the different types of properties. This represents objects as vectors of binary properties. While this approach is agnostic to downstream prediction tasks it may be biased since subjects may leave out important features and/or mention unimportant ones. For example, one may forego a general property (e.g., "is an animal") while providing a highly specific fact (e.g., "is found in Florida"). In an alternative, deductive approach, researchers postulate dimensions of interest and subsequently let subjects place objects in each dimension. Binder et al. [3] employed such an approach and collected ratings for hundreds of objects, verbs, and adjectives. These ratings were gathered over 65 dimensions, reflecting sensory, motor, spatial, temporal, affective, social, and cognitive experiences. Nonetheless, it is desirable to discover object representations that are not biased by the conducted behavioral task and whose dimensions are interpretable without necessitating *a priori* assumptions about their semantic content.

Recently, Zheng et al. [45] and Hebart et al. [18] introduced SPoSE, a model of the mental representations of 1,854 objects in a 49-dimensional space. The model was learned from human judgments about object similarity, where subjects were asked to determine an odd-one-out object in random triplets of objects. SPoSE embedded each object in a vector space so that each dimension is non-negative and sparse (most objects have a zero entry for a given dimension). The authors showed that the embedding dimensions of objects were interpretable and subjects could coherently label what the dimensions were "about," ranging from categorical (e.g., animate, food) to functional (e.g., tool),

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

structural (e.g., made of metal or wood), or visual (e.g., coarse pattern). The authors hypothesized that interpretability arose from combining positivity and sparsity constraints so that no object was represented by every dimension, and most dimensions were present for only a few objects. In addition SPoSE could predict human judgements close to the estimated best attainable performance [45, 18].

However, SPoSE has several limitations despite its notable performance. The first stems from the use of an ℓ_1 sparsity penalty to promote interpretability. In SPoSE, 6 to 11 dominant dimensions for an object account for most of the prediction performance. These dimensions are different between objects. A potential issue with enforcing SPoSE to have even fewer dimensions is that it may cause excessive shrinkage of the dominant values [1]. Second, when inspecting the distributions of values across objects, most SPoSE dimensions do not reflect the exponential prior induced by the ℓ_1 penalty. Overcoming this prior may lead to suboptimal performance and inconsistent solutions, specifically in low data regimes. Third, SPoSE uses an ad hoc criterion for determining the dimensionality of the solution via an arbitrary threshold on the density of each dimension. Finally, SPoSE has no criterion for determining convergence of its optimization process, nor does it provide any formal guarantees on the sample size needed to learn a model of desired complexity.

To overcome these limitations we introduce VICE, a variational inference (VI) method for embedding object concepts with interpretable, sparse, and positive dimensions. We start by discussing related work, then describe the triplet task and SPoSE before presenting theory and experimental results.

Contribution 1: VICE solves major limitations of SPoSE First, VICE encourages shrinkage while allowing for small entry values by using a *spike-and-slab* prior [5, 11, 14, 28, 35, 39, 43]. We deem this more appropriate than an exponential prior, because *importance* – the value an object takes in a dimension – is different from *relevance* – whether the dimension is applicable to that object – and they can be controlled separately. Second, we use VI with a unimodal posterior for representing each object in a dimension which yields a mean value and an uncertainty estimate. While unimodality makes it possible to use the mean values as representative object embeddings, the uncertainty estimates allow us to use a statistical procedure to automatically select the dimensions that explain the data. Third, we use this procedure to introduce a convergence criterion that reliably identifies *representational stability*, i.e., consistency in the number of selected dimensions.

Contribution 2: A PAC bound on the generalization of SPoSE and VICE models This bound can be used *retrospectively*, to provide guarantees about the generalization performance of a converged model. Furthermore it can be used *prospectively* to determine the sample size required to identify a representation, given the number of objects and a maximum possible number of dimensions.

Contribution 3: Extensive evaluation of model performance, across multiple datasets We compare VICE with SPoSE over three different datasets. One of these datasets contains concrete objects, another consists of adjectives, and the third is composed of food items. Experimentally we find that VICE meets or beats the performance of SPoSE in modeling human behavior. Moreover, we find that VICE performs substantially better regarding the reproducibility of its dimensions and displays a lower variance for the number of (selected) dimensions. Both of these measures are particularly important in the cognitive sciences. Lastly, we compare VICE with SPoSE for reduced amounts of data and show that VICE has significantly better performance on all measures.

A PyTorch implementation of VICE with Continuous Integration (CI) is available on GitHub (`https://anonymous.4open.science/r/VICE-59F0`) and will be de-anonymized upon acceptance.

2 Related work

Navarro & Griffiths [31] introduced a method for learning semantic concept embeddings from item similarity data which infers the number of embedding dimensions using the Indian Buffet Process (IBP) [17]. Their approach relies on continuous-valued similarity ratings rather than discrete forced-choice behavior and is not directly applicable to our setting. It is also challenging to scale the IBP to the number of dimensions and samples in our work [36]. Roads & Love [38] introduced a method for learning object embeddings from behavior in an 8-rank-2 task. Their method predicts behavior from

the embeddings by using active sampling [15] to query subjects with the most informative stimuli, yielding an object similarity matrix. Interpretability of embedding dimensions was not considered.

Other works have developed interpretable concept representations from text corpora. Early methods used word embeddings with positivity and sparsity constraints [30]. Later works in this direction used topic model representations of Wikipedia articles about objects [34], transformations of word embeddings into sparse non-negative representations [41, 32], or predictions of properties [9] or dimensions [44]. Others have considered using text corpora in conjunction with imaging data [12, 7]. Finally, Derby et al. [8] introduced a neural network function that maps the sparse feature space of a semantic property norm to the dense space of a word embedding, identifying informative combinations of properties, and ranking candidate properties for new words.

3 Methods

3.1 Triplet task

The *triplet task*, also known as the *odd-one-out* triplet task, is used for discovering object concept embeddings from similarity judgments over a set of m different objects. These judgments are collected from human participants who are given queries which consist of a *triplet* of objects (e.g., {"suit", "flamingo", "car"}). Participants are asked to consider the three pairs in a triplet {{"suit","flamingo"}, {"suit","car"}, {"flamingo","car")}, and to decide which pair is the most similar, leaving the third as the odd-one-out. We assign each object a numerical index, e.g., $1 \leftarrow$ "aardvark", $\ldots, 1854 \leftarrow$ "zucchini". Let $\{y, z\}$ denote the indices in this pair, e.g., $\{y, z\} = \{268, 609\}$ for "suit" and "flamingo." A dataset $\mathcal{D}$ is a set of n ordered pairs of presented triplets and selected pairs. That is, $\mathcal{D} := \big(\{i_s, j_s, k_s\}, \{y_s, z_s\}\big)_{s=1}^{n}$, where $\{y_s, z_s\} \subset \{i_s, j_s, k_s\}$.

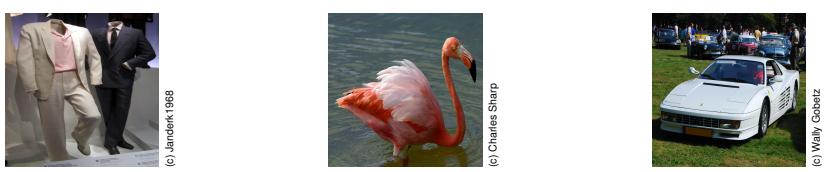

Figure 1: Example triplet consisting of suit-flamingo-car. (Creative Commons images)

3.2 SPoSE

Sparse Positive object Similarity Embedding (SPoSE) [45] is an approach for finding interpretable embedding dimensions from the triplet task. It does so by finding an embedding vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]$ for every object i. Let X denote the $m \times d$ matrix $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ and $S := XX^T$ be the similarity matrix, where S_{ij} denote its entry at i, j. The probability of choosing $\{y_s, z_s\}$ as the most similar pair of objects, given an object triplet $\{i_s, j_s, k_s\}$ and the embedding matrix X, is modeled as

$$p(\{y_s, z_s\}|\{i_s, j_s, k_s\}, X) := \frac{\exp(S_{y_s,z_s})}{\exp(S_{i_s,j_s}) + \exp(S_{i_s,k_s}) + \exp(S_{j_s,k_s})}. \tag{1}$$

At this point we remark that each triplet $\{i_s, j_s, k_s\}$ is sampled uniformly at random from the collection of all possible sets of object triplets $\mathcal{T}$. That is, $\{i_1, j_1, k_1\}, \ldots, \{i_n, j_n, k_n\} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\mathcal{T})$. This is precisely stated in A.1. SPoSE uses maximum a posteriori (MAP) estimation with a Laplace prior under a non-negativity constraint (equivalent to an exponential prior) to find the most likely embedding X for the training data $\mathcal{D}$. This leads to the training objective

$$\underset{X \geq 0}{\arg\min} \; -\log p(\mathcal{D}|X) + \lambda \sum_{i=1}^{m} \sum_{j=1}^{d} |X_{ij}|,$$

where $\log p(\mathcal{D}|X) = \text{constant} + \sum_{s=1}^{n} \log p(\{y_s, z_s\}|\{i_s, j_s, k_s\}, X)$ (see A.1 for derivation) and λ is determined using cross-validation.

3.3 VICE

In contrast to SPoSE, we use VI [4] instead of a MAP estimate for approximating the posterior probability $p(X|\mathcal{D})$. In VICE we impose additional constraints on the embedding matrix X by using a prior that encourages shrinkage while allowing entries in X to be close to zero (approx. sparse).

3.3.1 Variational Bayesian Inference

For VICE we consider approximating $p(X|\mathcal{D})$ with a variational distribution, $q_\theta(X)$, where $q_\theta \in \mathcal{Q}$, and θ is optimized to minimize the KL divergence to the true posterior, $p(X|\mathcal{D})$. In VICE the KL divergence objective function (derived in A.2) is

$$\arg\min_\theta \; \mathbb{E}_{q_\theta(X)} \left[\frac{1}{n} \left(\log q_\theta(X) - \log p(X) \right) - \frac{1}{n} \sum_{s=1}^{n} \log p\left(\{y_s, z_s\} | \{i_s, j_s, k_s\}, X \right) \right]. \quad (2)$$

Variational distribution VI requires a choice of a parametric variational distribution $q \in \mathcal{Q}$. For VICE we use a Gaussian distribution with a diagonal covariance matrix $q_\theta(X) = \mathcal{N}(\mu, \text{diag}(\sigma^2))$ where $\theta = \{\mu, \sigma\}$. Therefore, each embedding dimension has a *mean* and a *standard deviation*. We deem a Gaussian variational distribution appropriate for a variety of reasons. First, under certain conditions, the posterior is Gaussian in the infinite-data limit [23]. Second, a unimodal variational distribution makes it possible to use μ as representative object embeddings. A fixed representations is useful for downstream use cases that rely on a single embedding vector for each object. Use cases range from using the embeddings as targets in a regression task over unsupervised clustering of the embeddings to interpreting the embeddings. More complex variational families are not as practical for this. Third, a Gaussian posterior is a computationally convenient choice.

Similarly to Titsias & Lázaro-Gredilla [42], we use a Monte Carlo (MC) approximation of (2) by sampling a limited number of Xs from $q_\theta(X)$ during training. We generate X with the reparameterization trick [22, 37], $X_{\theta,\epsilon} = \mu + \sigma \odot \epsilon$, where $\epsilon \in \mathbb{R}^{m \times d}$ is entrywise $\mathcal{N}(0, 1)$, and $\odot$ denotes the Hadamard (element-wise) product. This leads to the objective

$$\arg\min_\theta \frac{1}{R} \sum_{r=1}^{R} \frac{1}{n} (\log q_\theta(X_{\theta,\epsilon^{(r)}}) - \log p(X_{\theta,\epsilon^{(r)}})) - \frac{1}{n} \sum_{s=1}^{n} \log p(\{y_s, z_s\} | \{i_s, j_s, k_s\}, [X_{\theta,\epsilon^{(r)}}]_+). \quad (3)$$

We apply a ReLU function, denoted by $[\cdot]_+$, to the sampled $X_{\theta,\epsilon}$ values to guarantee that $X_{\theta,\epsilon} \in \mathbb{R}_+^{m \times d}$. As commonly done in the Dropout and Bayesian Neural Network literature [40, 5, 13, 26], we set $R = 1$ for computational efficiency during the optimization process.

Posterior probability estimation Computational efficiency at inference time is not as critical as it is during training. Therefore, we can get a better estimate of the posterior probability distribution over the three possible odd one-one-out choices by letting $R \gg 1$. Using the optimized variational posterior, $q_{\hat{\theta}(X)}$, we approximate the probability distribution with an MC estimate [16, 5, 26, 4] from R samples $X^{(r)} = X_{\hat{\theta},\epsilon^{(r)}}$ for $r = 1, \ldots, R$, yielding

$$\hat{p}\left(\{y, z\} | \{i, j, k\} \right) \coloneqq \frac{1}{R} \sum_{r=1}^{R} p(\{y, z\} | \{i, j, k\}, X^{(r)}). \quad (4)$$

Spike-and-Slab prior As discussed above, SPoSE induces sparsity through an ℓ_1 penalty which, along with the non-negativity constraint, is equivalent to using an exponential prior. Through examination of the publicly available histograms of weight values in the two most important SPoSE dimensions (see Figure 2 in B.3), we observed that the dimensions did not resemble an exponential distribution. Instead, they contained a *spike* of probability at zero and a wide *slab* of probability for the non-zero values. To model this, we use a spike-and-slab Gaussian mixture prior [5, 11, 14, 20, 25],

$$p(X) = \prod_{i=1}^{m} \prod_{j=1}^{d} (\pi_{\text{spike}} \mathcal{N}(X_{ij} | 0, \sigma^2_{\text{spike}}) + (1 - \pi_{\text{spike}}) \mathcal{N}(X_{ij} | 0, \sigma^2_{\text{slab}})), \quad (5)$$

which encourages shrinkage. This prior has three parameters, σ_{spike}, σ_{slab}, and π_{spike}. π_{spike} is the probability that an embedding dimension is drawn from the *spike* Gaussian. Since spike and slab distributions are mathematically interchangeable, by convention we require that $\sigma_{\text{spike}} \ll \sigma_{\text{slab}}$.

3.3.2 Dimensionality reduction and convergence

For interpretability purposes it is desirable for the object embedding dimensionality, d, to be as small as possible. In contrast to SPoSE, which employs a user-defined threshold to prune dimensions, VICE exploits the uncertainty estimates for embedding values to select a subset of informative dimensions.

The pruning procedure works by assigning importance scores to each of the d dimensions, which reflect the number of objects that we can confidently say have a non-zero weight in a dimension. To compute the score, we use the variational embedding for each object i and dimension j – location μ_{ij} and scale σ_{ij} parameters – to compute the posterior probability that the weight is truncated to zero according to the left tail of a Gaussian distribution with that location and scale (see §3.3.1). This gives us a posterior probability of the weight being zero for each object within a dimension [16]. To calculate the overall importance of a dimension, we estimate the number of objects that have non-zero weights, while controlling False Discovery Rate (FDR) [2] with $\alpha = 0.05$. We define importance of each dimension j to be the number of objects for which $P(X_{ij} > 0) \geq .95$ holds. After convergence, we prune the model by removing dimensions with 5 or fewer statistically significant objects. This is a commonly used reliability threshold in semantic property norms (e.g., [27, 9]).

In gradient-based optimization, the gradient of an objective function with respect to the parameters of a model, $\nabla\mathcal{L}(\theta)$, is used to iteratively find parameters $\hat{\theta}$ that minimize that function. We use a *representational stability* criterion to determine convergence. That is, the optimization process halts when the number of identified dimensions - as described above - has not changed by a *single* dimension over the past L epochs (e.g., $L = 500$). Given that our goal is to find stable estimates of the number of dimensions, we considered this to be more appropriate than other convergence criteria such as evaluating the cross-entropy error on a validation set or evidence-based criteria [24, 10]. For further details on convergence and the optimization process see B.1.

3.4 Sample complexity bound

We use statistical learning theory to obtain estimates of the sample size needed to appropriately constrain VICE (and SPoSE) models. These estimates can be used *retrospectively*, to obtain probabilistic guarantees on the generalization of a trained model to unseen data, or *prospectively*, to decide how much data to collect for a study. For VICE, and in this analysis only, we extract μ and use it as a fixed embedding to predict the most likely odd-one-out for a query triplet rather than sampling embeddings from the variational distribution.

To obtain a useful bound we make two assumptions. First, we assume that there exists an upper bound M on the largest value in the embedding. Second, we assume that the embeddings obtained by either SPoSE or VICE can be quantized in a relatively coarse fashion with marginal losses in predictive performance. We can choose a certain discretization scale Δ, e.g., $\Delta = 0.5$, and round embedding values to a nonnegative integer multiple of Δ. While we employ this quantization mainly to use learning theory bounds for finite hypothesis classes, it could have benefits for interpretation as well (e.g., a dimension could consist of labels such as zero (0), very low (0.5), low (1), medium (1.5), high (2)). In all datasets we have used, SPoSE and VICE embeddings with reasonable priors had values below 2.7 (see E.2). Given an upper bound M and a discretization scale Δ one can consider all possible ways of populating the matrix X with weights limited to the set $S = \{0, \Delta, \ldots, (k-1)\Delta, k\Delta\}$. The following bound tells us that with high probability the estimated error rate and true error rate for such discretized X matrices are close to each other.

Proposition 3.1. *Given $\delta > 0$, $\epsilon > 0$, and $n \geq \left(md\log(k+1) + \log\left(1/\delta\right)\right) / \left(2\epsilon^2\right)$ training samples $\left(\{y_s, z_s\}, \{i_s, j_s, k_{1s}\}\right)_{s=1}^{n}$, then*

$$P\left(\sup_{X \in S^{m \times d}} \hat{R}(X) - R(X) < \epsilon\right) \geq 1 - \delta,$$

where

$$\hat{R}(X) := \frac{1}{n} \sum_{s=1}^{n} \mathbb{1} \left(\{y_s, z_s\} \neq \underset{\{y,z\}}{\arg\max}\, p\left(\{y,z\} | \{i_s, j_s, k_s\}, X\right) \right)$$

$$R(X) := P_{\{y',z'\}|\{i',j',k'\}} \left(\{y', z'\} \neq \underset{\{y,z\}}{\arg\max}\, p\left(\{y,z\} | \{i', j', k'\}, X\right) \right).$$

Proof sketch. This bound follows from applying Hoeffding's inequality to the $(k+1)^{m \times d}$ elements of $S^{m \times d}$ combined with a union bound. This proof is virtually identical to that of Theorem 2.13 in Mohri et al. [29]. A complete proof can be found in C. □

To use the bound we first decide on the number of quantization steps $k = \lceil M/\Delta \rceil$. The probability of violating the bound δ is analogous to the Type I error control in hypothesis testing, often 0.05. As the bound depends very weakly on δ, it is convenient to use $\delta = 0.001$. The error tolerance ϵ does have a major effect on the sample size estimate provided by the bound. For the bound to be practically useful, ϵ has to be smaller than the difference between the training error – average zero-one loss over the training set examples – and random guessing error. For the datasets we use in this paper we observed a difference of $0.2 - 0.3$. A conservative choice would halve the accuracy gap of 0.2, giving $\epsilon = 0.1$. Together, these values result in a prospective sample size estimate of $n \approx 50 \cdot (md \log(5) + \log(1000))$. To use the bound retrospectively, we fix n, δ while varying Δ to get a guarantee on ϵ, which in turn yields a probabilistic upper bound on the generalization error for an embedding, $R(X) \leq \hat{R}(X) + \epsilon$. For more details, see Algorithm 1 in E.1.

4 Experiments

4.1 Data

Our experiments were performed on three datasets: THINGS [45, 18], ADJECTIVES, and FOOD.[1] THINGS and ADJECTIVES contain random samples from all possible triplet combinations for 1,854 objects and 2,372 adjectives, respectively. THINGS and ADJECTIVES are each comprised of a large training dataset which contains no "repeats," i.e., there is only one human response for each triplet ($\sim$ 1.5M triplets for THINGS and $\sim$ 800K triplets for ADJECTIVES). For each of these training datasets 10% of the triplets are assigned to a predefined validation set. The test sets for both contain the results for 1,000 random triplets that are not contained in the training data, with 25 repeats for each triplet. FOOD contains data for every possible triplet combination for 36 objects and repeats for some triplets. We partitioned this dataset into train (45%), validation (5%), and test (50%) sets with disjoint triplets. F provides details about stimuli, data collection, and quality control.

4.2 Experimental setup

We implemented both SPoSE and VICE in PyTorch [33] using Adam [21] with $\eta = 0.001$. To find the optimal VICE hyperparameter combination we performed a grid search over $\sigma_{\text{spike}}, \sigma_{\text{slab}}, \pi_{\text{spike}}$, optimizing (3) and evaluating each model on the validation set. All experiments were performed over 20 different random initializations. For VICE we observed that the hyperparameter values with the lowest average cross-entropy error on the validation set were highly similar across the different datasets (see Table 1 in D). For more information about the experimental setup, including details about training, weight initialization(s), hyperparameter grid, and optimal combination, see D.

4.3 Evaluation measures

VICE estimates $p(\{y,z\}|\{i,j,k\})$, the model's softmax probability distribution over a given triplet (see (4)). The first evaluation measure we consider is the prediction *accuracy* with respect to

[1] ADJECTIVES and Food are not yet publicly available, but were shared with us by their authors.

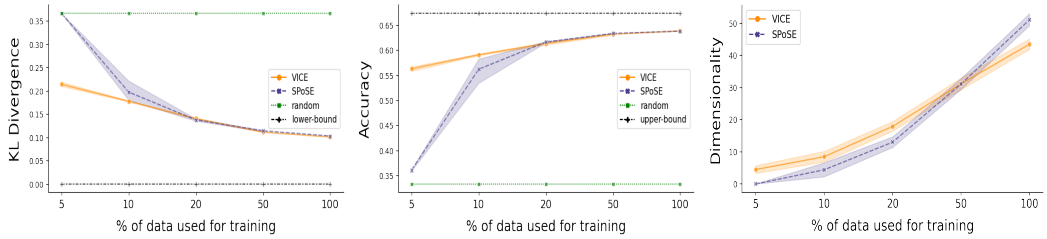

Figure 2: VICE vs. SPoSE: KL divergences between model output and human probability distributions (*left*), triplet task prediction accuracies (*center*), and number of identified embedding dimensions (*right*). VICE and SPoSE were each trained on differently sized subsets of the THINGS training data. Error bands depict 95% CIs across the different random initializations and partitions of the data.

the correct human choice $\{y, z\}$, where the predicted pair is $\arg\max_{\{y,z\}} \hat{p}\left(\{y, z\}|\{i, j, k\}\right)$. We estimate an upper bound on the prediction accuracy by using the repeats in the test set (see D). A Bayes-optimal classifier would predict the majority outcome for any triplet, whereas random guessing gives an accuracy of $1/3$. Note that the triplet task is subjective and thus there is not necessarily a definitive correct answer for a given triplet. If a test set contains multiple responses for a triplet, it provides information about the relative similarities of the three object pairs. This lets us calculate probability distributions over answers for each triplet. Approximating these distributions closely is important in cognitive science applications. Therefore, we additionally evaluate the models by computing the KL divergences between the models' posterior probability estimates $\hat{p}(\{y, z\}|\{i, j, k\})$ (see (4)) and the human probability distributions inferred from the test set across all test triplets and report the average.

4.4 Results on THINGS

Full dataset For this experiment we compared a representative model from VICE and SPoSE. The representative models were chosen to be those with the median cross-entropy error on the validation set over the 20 seeds. For VICE, we set the number of MC samples to $R = 50$. On the test set, VICE and SPoSE achieved similar prediction accuracies of 0.638 and 0.637 respectively (estimated upper bound for accuracy was 0.673). Likewise, VICE and SPoSE achieved similar average KL divergences of 0.100 and 0.103 respectively (random guessing KL divergence was 0.366). The differences between the median model accuracies and KL divergences were not statistically significant according to a two-sided paired *t*-test over individual test triplets. Hence, VICE and SPoSE predicted triplets approximately equally well when trained on the full dataset. This is not surprising, however, since Bayesian methods based on MC sampling become more like deterministic maximum likelihood estimation in the infinite data limit. If $n \to \infty$, then the left-hand side in the expectation of (2), which accounts for the contribution of the prior, goes to zero. Conversely, the effects of the prior are more prominent when n is small, as we show in the following section.

Data efficiency experiments Performance on small datasets is important in cognitive science, since behavioral experiments often have low sample sizes (e.g., tens to hundreds of volunteer in-lab subjects) compared to THINGS, which is particularly large, or they can be costly to crowdsource. To test whether VICE can model the data better than SPoSE in low sample regimes we performed experiments training on smaller subsets of the training dataset. We did this for subsets with sizes equal to 5%, 10%, 20%, and 50% of the training dataset. For each size we divided the training set into equal partitions and performed an experiment for each partition, thereby giving multiple results for each subset size. Validation and held-out test sets remained unchanged. In Figure 2 we show the average KL divergence (left) and prediction accuracy (middle) across data partitions and 20 random seeds (used to initialize model parameters) for various dataset sizes. The average performance across partitions was used to compute the confidence intervals (CIs). The differences in prediction accuracies and KL divergences between VICE and SPoSE were notable for the 5% and 10% data subsets. In the former, SPoSE predicted only slightly better than random guessing; in the latter, SPoSE showed

a large variability between random seeds and data partitions, as can be seen in the 95% CIs. In comparison VICE showed a small variation in the two performance metrics across random seeds and performed much better than random guessing. The differences between VICE and SPoSE for the 5% and 10% subset scenarios were statistically significant according to a two-sided paired *t*-test comparing individual triplet predictions between median models ($p < 0.001$). For the full training dataset VICE used significantly fewer dimensions than SPoSE to achieve the same performance.

4.5 Other results

Other datasets On the ADJECTIVES test set, VICE and SPoSE had accuracies of 0.559 and 0.562, respectively (estimated upper bound was 0.607), and KL divergences of 0.083 and 0.088, respectively (random guessing was 0.366). On the FOOD test set, VICE and SPoSE median models had accuracies of 0.693 and 0.698, respectively. The differences in accuracy (and KL divergence for ADJECTIVES) between VICE and SPoSE for both datasets were not statistically significant according to two-sided paired *t*-tests on the median model predictions across triplets on the test set. However, for FOOD, VICE used significantly fewer dimensions than SPoSE to achieve similar performance (see Table 1).

Hyperparameters We observed that VICE's performance is fairly insensitive to hyperparameter selection. Using default hyperparameters, $\sigma_{\text{spike}} = 0.25$, $\sigma_{\text{slab}} = 1.0$, $\pi_{\text{sigma}} = 0.5$, yields an accuracy score within 0.015 of the best cross-validated model in the full dataset setting, on all three datasets.

4.6 Reproducibility and Representational stability

Reproducibility As mentioned in § 1, a key criterion for learning concept representations beyond predictive performance is *reproducibility*, i.e., learning similar representations for different random initializations on the same training data. To evaluate this we compared the learned embeddings from 20 differently initialized VICE and SPoSE models. The first aspect of reproducibility we investigate is whether the models yield a consistent number of embedding dimensions across random seeds.

As reported in Table 1, VICE yields fewer dimensions than SPoSE with less variance in the number of dimensions for all three datasets. The embedding dimensionality was statistically significantly smaller according to an independent t-test for THINGS ($p < 0.001$) and FOOD ($p < 0.001$), but not for ADJECTIVES ($p = 0.108$). The difference in the standard deviations for ADJECTIVES was statistically significant according to a two-sided F-test ($p = 0.002$), but was not statistically significant for THINGS ($p = 0.283$) or FOOD ($p = 0.378$). In addition, we observed that the identified dimensionality is consistent as long as d is chosen to be sufficiently large (see D).

Table 1: Reproducibility of VICE and SPoSE. Reported are the means and standard deviations with respect to selected dimensions and the average reproducibility score of dimensions (in %) across random seeds. Bold means VICE performed statistically significantly better with $\alpha = 0.05$.

	VICE		SPoSE	
DATA\ METRIC	Selected Dims.	Reproducibility	Selected Dims.	Reproducibility
THINGS	**44** (1.59)	**87.01**%	52 (1.82)	81.30%
ADJECTIVES	21 (**0.77**)	76.76%	22 (1.53)	71.64%
FOOD	**5** (0.95)	**87.38**%	16 (1.02)	62.88%

The second aspect to reproducibility we examine is the degree to which the identified dimensions are similar across random initializations up to a permutation. To calculate the number of reproducible dimensions we associate each embedding dimension of a model with the most similar embedding dimension across the other models. We quantify reproducibility of a dimension as the average Pearson correlation between one dimension and its best match across the 19 remaining models. In Table 1 we report the average relative number of dimensions with a correlation > 0.8 across models. The embedding dimensions in VICE were more reproducible than those in SPoSE. The difference in average reproducibility scores was statistically significant according to a two-sided, independent t-test for FOOD ($p = 0.030$) and THINGS ($p = 0.008$), but not for ADJECTIVES ($p = 0.466$).

Representational stability For all three datasets VICE found reproducible and stable dimensions across all 20 random initializations and converged when the embedding dimensionality had not changed over the past $L = 500$ epochs. The median number of epochs to achieve representational stability was comparable for the similarly-sized THINGS and ADJECTIVES datasets, but occurred later for FOOD (likely due to the smaller number of gradient updates per epoch). In B.1 we show plots that demonstrate that the convergence criterion as defined in §3.3.2 worked reliably for all datasets.

4.7 Interpretability

Figure 3: Four example VICE dimensions showing the top six objects for THINGS (*top*) and FOOD (*middle*), and wordclouds for ADJECTIVES (*bottom*).

One of the benefits of SPoSE is the interpretability of the dimensions of its concept embeddings which was evidenced through experiments with humans [18]. VICE dimensions are equally interpretable due to similar constraints on its embedding space. Therefore, it is easy to sort objects within a dimension of the VICE mean embedding matrix μ in descending order, to obtain human judgments of what an embedding dimension represents. To illustrate the interpretability of VICE we show in Figure 3 four example dimensions of a representative model for THINGS, FOOD, and ADJECTIVES. For THINGS the dimensions appear to represent *categorical*, *functional*, *structural*, and *color-related* information. For FOOD the dimensions appear to be a combination of *fried*, *vegetable*, *fruit*, and *sweet* food items. For ADJECTIVES the dimensions seem to reflect *size/magnitude*, *visual appearance*, *negative valence*, and *sensory* adjectives. In G we show every VICE dimension for THINGS only, since manuscripts about the other datasets are still in preparation.

5 Conclusion

In this paper we introduced VICE, a novel VI approach for learning interpretable object concept embeddings by modeling human behavior in an odd-one-out triplet task. We showed that VICE predicts human behavior close to the estimated best attainable performance across three datasets and that VICE outperformed a competing method, SPoSE, in low sample regimes. In addition, VICE has several characteristics that are desirable for scientific use. It has an automated procedure for determining the number of dimensions needed to explain the data, which further enables the detection of convergence during training. This leads to better model reproducibility across different random initializations and hyperparameter settings. As a result, VICE can be used out-of-the-box without the necessity to perform an extensive search over random seeds or tuning model hyperparameters. Finally, we introduced a PAC learning bound on the generalization performance for a VICE model. One of the limitations of VICE is that it assumes a shared mental representation across participants, akin to much of the literature. However, the VI framework can be leveraged to model inter-individual differences, which we plan to do in future work.

References

[1] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. doi: 10.3150/11-BEJ410.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: https://doi.org/10.1111/j.2517-6161.1995.tb02031.

[3] Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4):130–174, 2016.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1613–1622. JMLR.org, 2015.

[6] Erin M. Buchanan, K. D. Valentine, and Nicholas P. Maxwell. English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, 51(4): 1849–1863, May 2019. doi: 10.3758/s13428-019-01243-z.

[7] Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. In Anna Korhonen and Ivan Titov (eds.), *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pp. 260–270. Association for Computational Linguistics, 2018. doi: 10.18653/v1/k18-1026.

[8] Steven Derby, Paul Miller, and Barry Devereux. Feature2vec: Distributional semantic modelling of human property knowledge. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 5852–5858. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1595.

[9] 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, December 2013. doi: 10.3758/s13428-013-0420-4.

[10] David Duvenaud, Dougal Maclaurin, and Ryan Adams. Early stopping as nonparametric variational inference. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1070–1077, Cadiz, Spain, 09–11 May 2016. PMLR.

[11] Ludwig Fahrmeir, Thomas Kneib, and Susanne Konrath. Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20(2):203–219, 2010. doi: 10.1007/s11222-009-9158-3.

[12] Alona Fyshe, Partha Pratim Talukdar, Brian Murphy, and Tom M. Mitchell. Interpretable semantic vectors from a joint model of brain- and text-based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 489–499. The Association for Computer Linguistics, 2014. doi: 10.3115/v1/p14-1046.

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1050–1059. JMLR.org, 2016.

[14] Edward I. George and Robert E. McCulloch. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. doi: 10.1080/01621459.1993.10476353.

[15] Jacqueline Gottlieb and Pierre-Yves Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):758–770, November 2018. doi: 10.1038/s41583-018-0078-0.

[16] Alex Graves. Practical variational inference for neural networks. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 2348–2356, 2011.

[17] Thomas L. Griffiths and Zoubin Ghahramani. The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

[18] Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multi-dimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, October 2020. doi: 10.1038/s41562-020-00951-3.

[19] Mariam Hovhannisyan, Alex Clarke, Benjamin R. Geib, Rosalie Cicchinelli, Zachary Monge, Tory Worth, Amanda Szymanski, Roberto Cabeza, and Simon W. Davis. The visual and semantic features that predict object memory: Concept property norms for 1, 000 object images. *Memory & Cognition*, 49(4):712–731, January 2021. doi: 10.3758/s13421-020-01130-5.

[20] Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[22] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[23] Bas JK Kleijn and Aad W van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

[24] Maren Mahsereci, Lukas Balles, Christoph Lassner, and Philipp Hennig. Early stopping without a validation set. *arXiv e-prints*, art. arXiv:1703.09580, March 2017.

[25] Gertraud Malsiner-Walli and Helga Wagner. Comparing spike and slab priors for bayesian variable selection. *arXiv e-prints*, art. arXiv:1812.07259, December 2018.

[26] Patrick McClure and Nikolaus Kriegeskorte. Robustly representing uncertainty in deep neural networks through sampling. *arXiv e-prints*, art. arXiv:1611.01639, November 2016.

[27] 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, November 2005. doi: 10.3758/bf03192726.

[28] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. doi: 10.1080/01621459.1988.10478694.

[29] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning Series. The MIT Press, 2018.

[30] Brian Murphy, Partha Talukdar, and Tom Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING 2012*, pp. 1933–1950, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[31] Daniel J Navarro and Thomas L Griffiths. Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, 20(11):2597–2628, 2008.

[32] Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. Word2sense: Sparse interpretable word embeddings. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5692–5705. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1570.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019.

[34] Francisco Pereira, Matthew Botvinick, and Greg Detre. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence*, 194:240–252, 2013. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.06.005. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

[35] Kolyan Ray, Botond Szabó, and Gabriel Clara. Spike and slab variational Bayes for high dimensional logistic regression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[36] Colorado Reed and Zoubin Ghahramani. Scaling the Indian Buffet Process via Submodular Maximization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 1013–1021. JMLR.org, 2013.

[37] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

[38] Brett D Roads and Bradley C Love. Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3547–3557, 2021.

[39] Veronika Rocková and Nicholas Polson. Posterior concentration for sparse deep learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual*

Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp. 938–949, 2018.

[40] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[41] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard H. Hovy. SPINE: sparse interpretable neural embeddings. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 4921–4928. AAAI Press, 2018.

[42] Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1971–1979, Bejing, China, 22–24 Jun 2014. PMLR.

[43] Michalis K. Titsias and Miguel Lázaro-Gredilla. Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24: Annual Conference on Neural Information Processing Systems 2011, NeurIPS 2011, December 12-14, 2011, Granada, Spain*, pp. 2339–2347, 2011.

[44] Akira Utsumi. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6), 2020. doi: 10.1111/cogs.12844.

[45] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See § 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See `https://anonymous.4open.science/r/VICE-59F0` for our codebase. See `https://osf.io/jum2f/` and `https://things-initiative.org/` for data.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See § 4.1, § 4.2 and D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 2 and Table 1.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See D
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See `https://anonymous.4open.science/r/VICE-59F0`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See F.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

   We re-used crowdsourced data collected by various collaborators, and attribute it to them in the main paper. Even though we did not collect these data ourselves, we provide a summary of the information asked for in 5. in F, as well as a short description of the task and stimuli in Section 3.1.