Predicting Human Similarity Judgments Using Large Language Models

Raja Marjieh [*1] **Ilia Sucholutsky** [*2] **Theodore R. Sumers** [2] **Nori Jacoby** [3] **Thomas L. Griffiths** [1 2]

Abstract

Similarity judgments provide a well-established method for accessing mental representations, with applications in psychology, neuroscience and machine learning. However, collecting similarity judgments can be prohibitively expensive for naturalistic datasets as the number of comparisons grows quadratically in the number of stimuli. We leverage recent advances in language models and online recruitment, proposing an efficient domain-general procedure for predicting human similarity judgments based on text descriptions. Crucially, the number of descriptions required grows only linearly with the number of stimuli, drastically reducing the amount of data required. We test this procedure on six datasets of naturalistic images and show that our models outperform previous approaches based on visual information.

1. Introduction

Similarity judgments are at the heart of the study of mental representations in the cognitive sciences as exemplified by the method of multi-dimensional scaling (MDS) (Shepard, 1980) as well as by a large corpus of work that followed (Shepard, 1987; Ghirlanda & Enquist, 2003; Battleday et al., 2020; Peterson et al., 2018; Jha et al., 2020; Caplette & Turk-Browne, 2022; Hebart et al., 2020). Moreover, similarity judgments play an important role in other disciplines such as neuroscience, e.g., in the method of representational similarity analysis (Kriegeskorte et al., 2008), as well as in machine learning, e.g., as a way to regularize latent spaces so that they align with human representations and perception (Esling et al., 2018).

Despite the success of similarity-based approaches, their reliance on pairwise comparisons that scale quadratically in the number of stimuli poses a serious limitation on their scalability. To reduce this burden, we leverage the deep relationship between conceptual structure and language (Murphy, 2002) to use linguistic descriptions as a proxy for human semantic representations. Intuitively, stimuli that are judged to be highly similar are likely to evoke similar descriptions, allowing us to use description similarity to predict pairwise similarity judgments. This approach offers two key advantages over prior work: first, it is *scalable*. While pairwise similarity comparisons scale quadratically with the number of stimuli (Shepard, 1980), text descriptions scale linearly. Second, it is *domain-general*: unlike CNN representations (Peterson et al., 2018), which are limited to visual stimuli, our procedure could be applied to any domain.

Finally, we note that our approach leverages two distinct and important advances. First, text descriptions can be easily crowd-sourced via online recruitment platforms such as Amazon Mechanical Turk (AMT; `https://www.mturk.com/`) and are part of the common practice in modern machine learning pipelines (Parekh et al., 2020). Second, modern language models (Speer et al., 2017; Devlin et al., 2018; Gao et al., 2021) provide rich latent representations of text. It is therefore natural to ask: how far can we go in predicting human similarity judgments based on language alone?

We explore this question on a collection of six datasets of naturalistic images for which the ground-truth similarity matrices are known (Peterson et al., 2018). Our exploration proceeds in three stages. In Study 1, we construct similarity estimates by applying a state-of-the-art word embedding model known as ConceptNet NumberBatch (CNNB) (Speer et al., 2017) to pre-existing semantic labels for the dataset images. In Study 2, we generalize this approach by crowd-sourcing free text descriptions from AMT, then constructing similarity estimates based on a variant of BERT (Devlin et al., 2018) tuned for semantic representations known as SimCSE (Gao et al., 2021). Finally, we combine the concept-level representation of CNNB with the fine-grained textual representation of SimCSE and generate a joint predictor of similarity judgments. In the process, we benchmark our models' predictive accuracy against the CNN-based approach of Peterson et al. (2018).

[*]Equal contribution [1]Department of Psychology, Princeton University, NJ, USA [2]Department of Computer Science, Princeton University, NJ, USA [3]Computational Auditory Perception Group, Max Planck Institute for Empirical Aesthetics. Correspondence to: Raja Marjieh <raja.marjieh@princeton.edu>, Ilia Sucholutsky <is2961@princeton.edy>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

2. General Methodology

Our general pipeline consists of collecting or using pre-existing linguistic descriptors for the individual stimuli and then using an embedding model to compute a proxy for pairwise similarity (Figure 1).

2.1. Predicting Human Similarity

Given a set of stimuli and their linguistic descriptors (semantic labels or free-text descriptions) as well as a suitable embedding scheme (e.g., a word embedding model) we used cosine similarity between the vectors representing two stimuli as the metric for calculating their similarity. Peterson et al. (2018) showed that predicting human similarity using CNN representations can be substantially enhanced by linearly transforming those representations. Mathematically, this corresponds to substituting the dot product $\mathbf{z}_1^T\mathbf{z}_2$ with $\mathbf{z}_1^T\mathbf{W}\mathbf{z}_2$ where $\mathbf{W}$ is a suitable diagonal matrix and $\mathbf{z}_1$ and $\mathbf{z}_2$ are the embedding vectors. Moreover, Peterson et al. showed that such a transformation can be found using ridge regression with L2 normalization. We apply this approach to our linguistic representations, using the Python library scikit-learn's RidgeRegression and RidgeCV implementations. To avoid overfitting and simulate generalization in practice, we performed 6-fold cross-validation over images which ensured that no images from the training set are present in the validation set. This ensures that even when combining SimCSE and CNNB representations, where the number of features increases, overfitting is still avoided. To facilitate comparison with previous work we quantified performance by computing Pearson R^2 scores (variance explained) (Peterson et al., 2018; Jha et al., 2020).

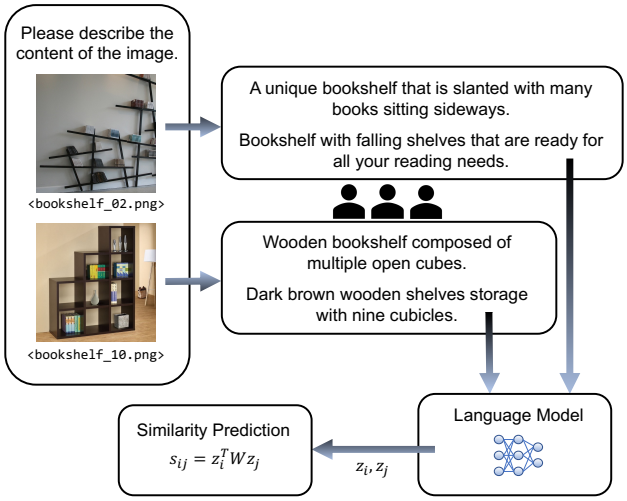

Figure 1. Schematic of the similarity prediction procedure based on text descriptions.

2.2. Stimuli

The six image datasets used in this paper were taken from Peterson et al. (2018). The datasets were organized based on six broad categories, namely, animals, fruits, vegetables, automobiles, furniture and various objects, each comprising 120 unique images. For all categories except animals, the datasets included semantic labels for each of the individual images. In the case of animals, we manually labeled the images. Sample images and labels appear in Figure 2.

3. Predicting Human Similarity Based on Semantic Labels

To initiate our investigation we first considered using the pre-existing semantic labels for the images in our datasets, as they served as concise summaries of the content of the images. We evaluated two representations for predicting human similarity judgments based on these labels: a one-hot representation and a word embedding representation.

3.1. One-hot Label Representation

The first approach served as a baseline and consisted of using the semantic labels as class labels with a "one-hot" representation.This representation implies that images with the same semantic label are maximally similar whereas images with different semantic labels are maximally dissimilar.

Figure 2. Sample images and their semantic labels.

This simple representation possessed non-trivial predictive power, as indicated by its average raw (i.e. before linear transformation) R^2 score of 0.31[1] across the datasets (see Appendix for all scores).

Finding positive but not strong correlations is not surprising as the one-hot representation misses fine-grained similarity between related (though not identical) semantic labels. Indeed, although a tiger and a leopard are distinct animals,

[1]The sparsity of one-hot representations makes linear transformation ineffective. To remedy this, we applied label smoothing ($\epsilon = 0.8$, see Appendix) to all the one-hot vectors. Linear transformation then resulted in a small boost in performance ($R^2 = 0.40$).

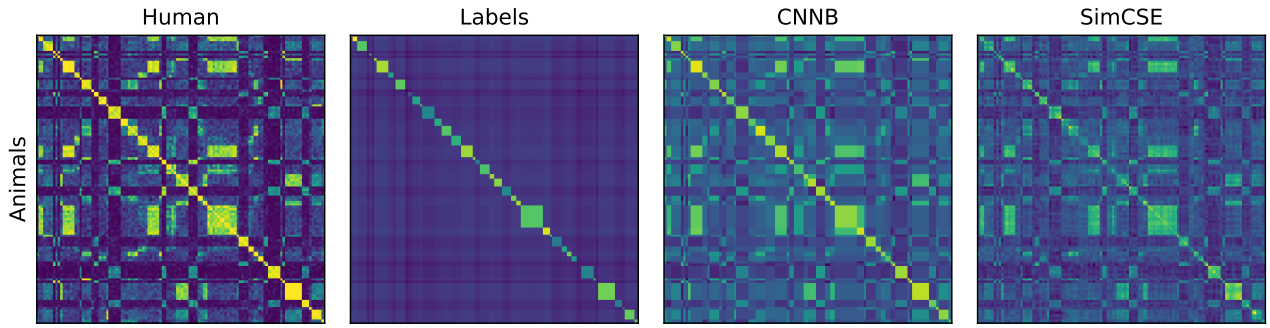

Figure 3. Full similarity matrices for the "animals" dataset for human participants (left), with corresponding predictions based on class labels, CNNB and SimCSE representations.

Table 1. R^2 scores on held-out images for linearly transformed representations from the different models.

Model	Animals	Automobiles	Fruits	Vegetables	Furniture	Various	$\langle R^2 \rangle$
CNNB	0.72	**0.86**	0.38	0.43	**0.63**	0.73	0.63
SimCSE	0.64	0.55	0.33	0.39	0.49	0.64	0.51
CNNB + SimCSE	**0.78**	0.83	**0.47**	**0.55**	0.62	**0.76**	**0.67**
CNN*	0.74	0.58	0.36	0.35	0.35	0.54	0.49

Note: $\langle R^2 \rangle$ is the average across all datasets. *CNN values are from Peterson et al. (2018). Additional results are in the appendix.

they nevertheless share some intuitive semantic similarity being members of the cat family; likewise for a chair and a recliner, or a strawberry and a blackberry. This can be seen in the absence of off-diagonal structure in the predicted similarity matrix (Figure 3). Nevertheless, this preliminary study serves as an initial evidence for the fact that people's judgments are indeed driven by semantic similarity.

3.2. Word-embedding Representation

To capture the structure of similarity between different semantic labels we replaced the one-hot representation with the latent representation of a state-of-art word embedding model known as ConceptNet NumberBatch (CNNB). CNNB is pre-trained on the ConceptNet knowledge graph (`https://conceptnet.io/`) which is targeted at capturing intuitive commonsense conceptual relations.

The use of CNNB representations resulted in a substantial performance boost over one-hot representations.The predicted similarity matrix is shown in Figure 3 and it is clear that a substantial part of the off-diagonal structure is recovered. To ensure that the linear transformation is not overfitting the similarity matrices, we performed 6-fold cross-validation as mentioned above and computed a control cross-validated (CCV) R^2 score on held-out images. These scores remained high ($R^2 = 0.63$), outperforming the CNN model of Peterson et al. (2018) (Table 1) on almost all datasets (except Animals, where it scored lower by a small margin). This implies that CNNB representations generalize better to new data. We also note that the dimensionality of the latent space of CNNB ($d = 300$) is much lower than that of the CNN ($d = 4096$) reducing the number of possible parameters to optimize over and hence the risk of overfitting.

4. Predicting Human Similarity Based on Free Text Descriptions

Concise semantic labels (and corresponding embeddings) are not always available for stimuli of interest. A more general approach would rely on free-text descriptions, which can be easily crowd-sourced online. Such data, however, requires a different kind of representations capable of flexibly encoding entire sentences (as opposed to aggregating representations of individual words which could lose important within-sentence structure). To that end, we used the latent representations of SimCSE (Gao et al., 2021) to embed free-text descriptions for each of the individual images which we

crowd-sourced on AMT[2]. The data collection procedure as well as example text descriptions are shown in Figure 1 (see Appendix for details on the embeddings).

4.1. Results

We used the embeddings to produce similarity estimates as before. We found that while the raw representations of SimCSE did not constitute a strong predictor, the linearly reweighted SimCSE representations ($d = 768$) demonstrated generalization performance comparable to the CNN-based model ($d = 4096$) of Peterson et al. (2018) (Table 1), though not as high as CNNB. One possible explanation for this difference is that CNNB predictors used single concise labels per image whereas for SimCSE we averaged representations of multiple descriptions which could capture different aspects of the image (Parekh et al., 2020). A more sophisticated approach could learn to pool embeddings from different descriptions efficiently; however for the purpose of the current work we chose to focus on simple linear transformations.

As a last step, we constructed a combined predictor that stacked CNNB and SimCSE representations to capture broad concept-level knowledge as well as fine-grained descriptions. The combined model resulted in the best aggregated performance, improving further on the CNNB model (Table 1).

5. Discussion

We proposed a highly efficient and domain-general procedure for predicting human similarity judgments based on text descriptions with linear (as opposed to quadratic) complexity. We tested our approach on six datasets of naturalistic images, finding evidence for its validity as well as outperforming previous models that rely on CNNs. These results suggest that human similarity judgments are indeed grounded in semantic understanding and language. Our work also provides a new perspective on the representational similarity between BERT and humans (Lake & Murphy, 2021).

In addition to psychological applications, our paradigm may allow for advances in machine learning. First, enriching machine learning datasets with similarity judgments and behavioral data more generally can endow artificial models with a variety of useful properties, such as human alignment and robustness against adversarial attacks (Peterson et al., 2019). Collecting similarity judgments over all pairs is infeasible for such datasets due to the large number of stimuli. Nevertheless, in many real-life applications similarity matrices tend to be sparse, i.e. only a small subset of comparisons would yield non-vanishing similarity (Parekh et al., 2020). An efficient enrichment pipeline, therefore, must exploit this sparsity and our procedure is a promising candidate for guiding such methods by predicting which pairs are likely to be informative *a priori*. Second, a systematic study comparing semantic and visual representations such as those of CNN and CNNB to the ones arising from human similarity could shed light on critical divergences between human and machine representations (Figure 4). We hope to engage with all of these avenues in future research.

[2]We also tried vanilla BERT but SimCSE outperformed it.

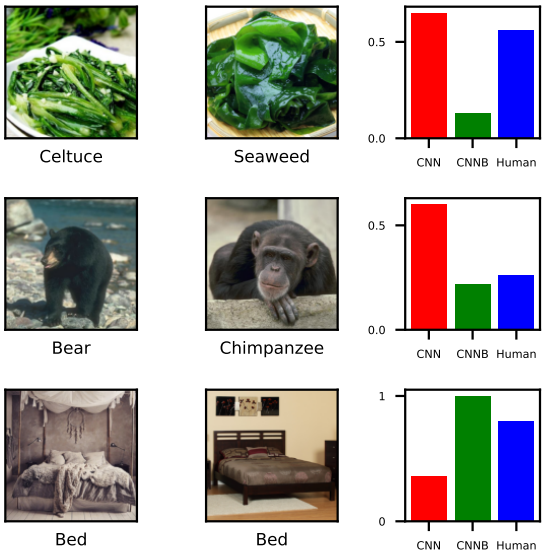

Figure 4. Examples of image pairs that generated large discrepancies between CNN and CNNB model predictions and their relation to human similarity scores.

References

Battleday, R. M., Peterson, J. C., and Griffiths, T. L. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):1–14, 2020.

Caplette, L. and Turk-Browne, N. Computational reconstruction of mental representations using human behavior. *PsyArxiv*, 2022. doi: https://doi.org/10.31234/osf.io/7fdvw.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Esling, P., Bitton, A., et al. Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics. *arXiv preprint arXiv:1805.08501*, 2018.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Ghirlanda, S. and Enquist, M. A century of generalization. *Animal Behaviour*, 66(1):15–36, 2003. ISSN 0003-3472. doi: https://doi.org/10.1006/anbe.2003.2174.

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., ..., and Jacoby, N. Gibbs sampling with people. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10659–10671. Curran Associates, Inc., 2020.

Hebart, M. N., Zheng, C. Y., Pereira, F., and Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

Jha, A., Peterson, J., and Griffiths, T. L. Extracting low-dimensional psychological representations from convolutional neural networks. *arXiv preprint arXiv:2005.14363*, 2020.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

Lake, B. M. and Murphy, G. L. Word meaning in minds and machines. *Psychological review*, 2021.

Lemhöfer, K. and Broersma, M. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, 44(2):325–343, 2012.

Murphy, G. *The big book of concepts*. MIT Press, 2002.

Parekh, Z., Baldridge, J., Cer, D., Waters, A., and Yang, Y. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. *arXiv preprint arXiv:2004.15020*, 2020.

Peterson, J. C., Abbott, J. T., and Griffiths, T. L. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669, 2018.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9617–9626, 2019.

Shepard, R. N. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980. doi: 10.1126/science.210.4468.390.

Shepard, R. N. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987. doi: 10.1126/science.3629243.

Speer, R., Chin, J., and Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020.

A. Crowdsourcing Methodology

The recruitment and experimental pipeline were automated using PsyNet (Harrison et al., 2020), a framework for experimental design which builds on top of the Dallinger platform[3] for recruitment automation. Overall, 328 US participants completed the study and they were paid $12 per hour. Upon completing a consent form participants had to take a standardized LexTALE English proficiency test (Lemhöfer & Broersma, 2012) to ensure caption quality. Participants that failed the pre-screening test were excluded from the study. Next, participants received the following instructions: "In this experiment we are studying how people describe images. You will be presented with different images and your task will be to describe their content. In doing so, please keep in mind the following instructions, 1) describe all the important parts of the image, 2) do not start the sentences with "There is" or "There are", 3) do not describe unimportant details, 4) you are not allowed to copy and paste descriptions, 5) descriptions should contain at least 5 words, 6) descriptions should contain at least 4 unique words. Note: no prior expertise is required to complete this task, just describe what you intuitively think is important as accurately as possible." Participants were then presented with nine random images from the dataset to help give them a sense of the images they were about to describe.

In each trial of the main experiment participants saw one of the images along with the following prompt "Please describe the content of the following image" (no semantic labels were provided). They then provided their description in a free text response box, subject to the constraints listed above. Each participant provided up to 30 text descriptions with each image receiving 15 text descriptions on average. To ensure that participants did not provide repetitive responses we computed the average Levenshtein edit distance between their current response and all previous responses. Participants for whom the average distance was close to zero (< 0.2) after 5 trials were excluded from the study. Any remaining random or very poor quality strings were excluded in a post-processing stage.

B. Pre-processing

Label Smoothing If $\vec{v}$ is the one-hot vector, then $\vec{v}_{smooth} = (1-\epsilon)\vec{v} + \frac{\epsilon}{k-1}(1-\vec{v})$ where ϵ is the smoothing parameter (we use a value of 0.8) and k is the number of classes (which is equal to the length of the vector). Smoothing does not change the relative structure of the resulting matrix but allows linear transformation to be successfully applied to the new vectors.

CNNB Embeddings CNNB contains embeddings not only for single words but also concepts consisting of several words. To make use of these, labels consisting of multiple words needed to have spaces replaced by underscores (e.g. 'red onion' becomes 'red_onion'). In addition, while the CNNB dictionary is quite large, there are certain words or concepts that it does not contain. In some of these cases, labels consisting of multiple words whose joint form was not found in CNNB had to be separated into individual words and their joint embedding estimated by their normalized sum (e.g. CNNB(animal body) $\approx \frac{\text{CNNB(animal)+CNNB(body)}}{\sqrt{2}}$). In other cases, labels had to be replaced by a synonym or the closest matching concept available in CNNB (e.g. 'tatsoi' was replaced by 'spoon_mustard').

Computing Semantic Embeddings We used a pre-trained SimCSE model, `sup-simcse-bert-base-uncased`, (Gao et al., 2021), accessed via the HuggingFace library (Wolf et al., 2020).[4] We extracted the semantic embeddings for each description, then took the average embedding across all descriptions for each image. In order to combine the SimCSE and CNNB representations, we first normalized both sets of embeddings by their respective means and standard deviations, and then concatenated the SimCSE and CNNB embeddings to get a single vector for each image.

C. MDS

To appreciate the semantic content of the predicted similarity matrices, we computed two-dimensional MDS representations of the images using the scikit-learn library with a maximum iteration limit of $10,000$ and a convergence tolerance of 1e-100. First metric MDS was applied to get an initial embedding, then four iterations of non-metric MDS were applied and the best solution was picked. The results are shown in Figure 5, and reveal a rich, interpretable semantic organization of the stimuli capturing a variety of semantic dimensions such as natural and functional classes and color gradients.

D. Additional Results

Table 2 contains all results of the four competing models. "Raw" corresponds to raw representations. "LT-Train" corresponds to linearly transformed representations evaluated on the training set. "LT-CCV" corresponds to linearly transformed representations evaluated on held-out images. CNN values are reproduced from Peterson et al. (2018).

[3] https://github.com/Dallinger/Dallinger

[4] We also used `bert-base-uncased`, but found that SimCSE outperformed it.

Table 2. R^2 scores for the different prediction models and datasets.

Model	Methodology	Animals	Automobiles	Fruits	Vegetables	Furniture	Various	$\langle R^2 \rangle$
Labels	Raw	0.23	0.69	0.20	0.24	0.34	0.19	0.31
CNNB	Raw	0.51	0.64	0.17	0.17	0.31	0.29	0.35
SimCSE	Raw	0.42	0.41	0.14	0.19	0.32	0.35	0.31
CNN*	Raw	0.58	0.51	0.27	0.19	0.37	0.27	0.37
Labels	LT-Train	0.29	0.71	0.26	0.27	0.38	0.48	0.40
CNNB	LT-Train	0.85	0.86	0.53	0.60	0.67	0.72	0.71
SimCSE	LT-Train	0.84	0.75	0.59	0.63	0.63	0.80	0.71
CNN*	LT-Train	0.84	0.79	0.53	0.67	0.72	0.52	0.68
CNNB	**LT-CCV**	**0.72**	**0.86**	**0.38**	**0.43**	**0.63**	**0.73**	**0.63**
SimCSE	**LT-CCV**	**0.64**	**0.55**	**0.33**	**0.39**	**0.49**	**0.64**	**0.51**
CNNB + SimCSE	**LT-CCV**	**0.78**	**0.83**	**0.47**	**0.55**	**0.62**	**0.76**	**0.67**
CNN*	**LT-CCV**	**0.74**	**0.58**	**0.36**	**0.35**	**0.35**	**0.54**	**0.49**

Note: "Raw" corresponds to raw representations. "LT-Train" corresponds to linearly transformed representations evaluated on training set, "LT-CCV" corresponds to linearly transformed representations evaluated on held-out images. $\langle R^2 \rangle$ is the average across all datasets. *CNN values are from Peterson et al. (2018).

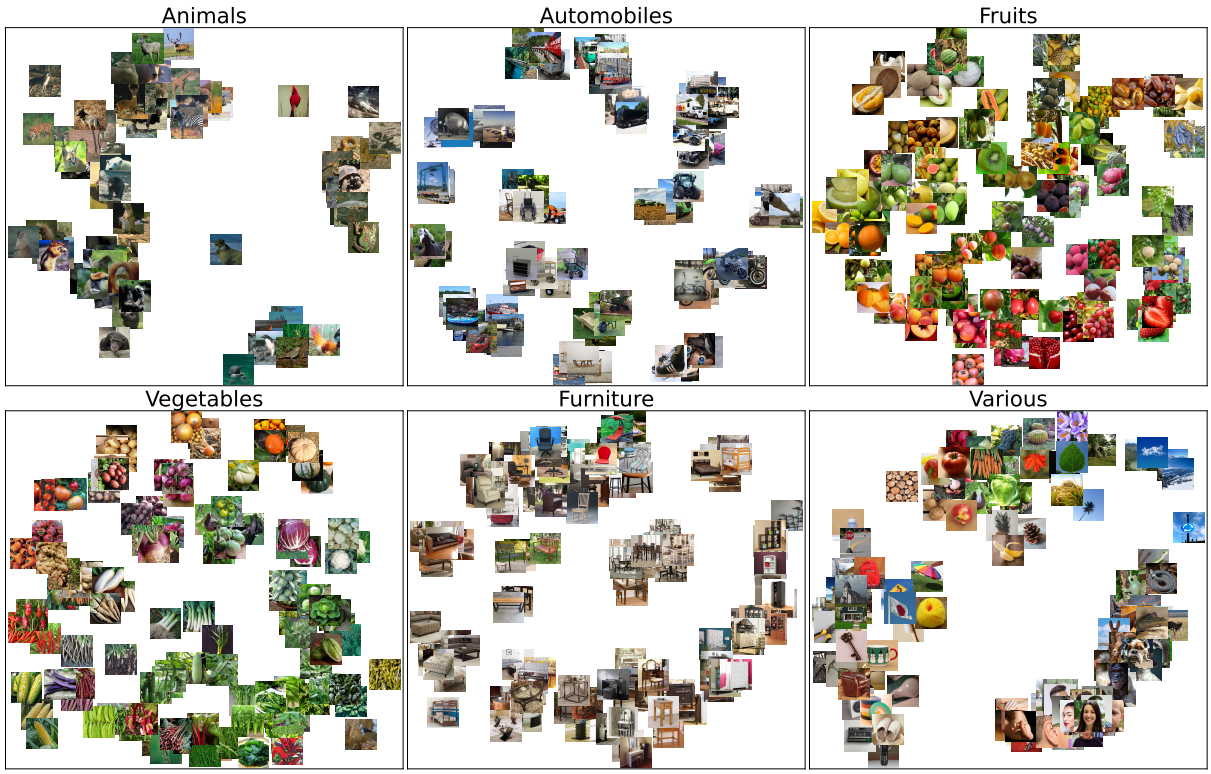

Figure 5. Two-dimensional MDS embedding of the joint CNNB-SimCSE similarity predictions.