Revealing the Feature Dimensions Driving Similarity Judgements of Natural Scenes

Peter Brotherwood (peter.brotherwood@umontreal.ca)[1]
Andrey Barsky (a.barsky@bham.ac.uk)[2]
Jasper van den Bosch (vandejjf@bham.ac.uk)[2]
Kendrick Kay (kay@umn.edu)[3]
Ian Charest (ian.charest.1@umontreal.ca)[1,2]

1 - cerebr**UM**, Département de Psychologie, Université de Montréal, Montréal, Quebec, Canada
2 - CHBH, School of Psychology, University of Birmingham, Birmingham, England, United Kingdom
3 - CMRR, Department of Radiology, University of Minnesota, Minneapolis, Minnesota, United States of America

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0

Abstract

Revealing the feature dimensions relevant in distinguishing and recognising natural scenes holds the key to understanding those features most important in driving our ability to perceive, recognise and adapt to changes in our surrounding environment. Previous work revealed the feature dimensions utilised in driving similarity judgements of a large set of stimuli from the THINGS dataset using a triplet odd-one-out similarity task. Here, we developed a similar approach to reveal feature dimensions relevant in driving similarity judgements from individual participants and stimuli from the Natural Scenes Dataset (NSD). Using similarity judgments from a multiple arrangements method, we train sparse positive similarity embeddings to predict pairwise similarities over a set of 100 diverse natural scenes. The resulting embeddings are capable of reproducing similarity judgements for each subject with a high degree of accuracy, with an average Spearman's correlation between embedding-predicted RDMs and RDMs produced by traditional iterative weighted averaging of similarity judgements of 0.88. Inspection of the resulting embeddings reveals the key feature dimensions driving these similarity judgements.

Keywords: Feature Space; Natural Scenes; Scene Recognition; Multiple Arrangements; SPosE

Introduction

Our ability to distinguish and recognise stimuli has long been thought to rely on the level of similarity between them (Ashby & Perrin, 1988; Ashby & Townsend, 1986). Revealing the features most commonly utilised in driving similarity judgements of stimuli would therefore help us identify which features drive our overall perception and recognition of visual scenes. Hebart et al. (2020) previously revealed common feature dimensions used in discriminating between simple objects from the THINGS dataset (Hebart et al., 2019) using similarity judgements from a large group of participants performing a triplet odd-one-out task. While useful for revealing the core feature dimensions that are shared across people, the number of data points required to estimate similarities using a triplet odd-one-out task makes it challenging to investigate feature spaces in individual participants for even moderately sized stimulus sets. To circumvent this difficulty, we developed a similar embedding strategy applied to similarity judgements collected using a multiple arrangements (MA) procedure (Kriegeskorte & Mur, 2012). The MA procedure is efficient at capturing the multidimensional representational space of similarities in individual participants, and these predict the unique nature of brain representations (Charest et al., 2014), but it does not allow the comprehensive understanding of the underlying dimensions offered with the technique developed by Hebart et al. (2020). Therefore, developing an approach combining the sparse positive similarity embeddings with data collected using the MA method offers the advantage of revealing rich feature dimensions capable of explaining individual differences in visual representations. Here, we present a hybrid approach where MA data was modelled using embeddings for a set of 100 natural scene stimuli in eight different individuals, measured in the Natural Scenes Dataset (NSD) (Allen et al., 2022).

Methods

Multiple Arrangements method

The 100 natural scenes were chosen such that their caption-level representations (as indexed using a deep averaging network; (Cer et al., 2018)) summarised the richness of the semantic representational space of NSD stimuli. The judgments were acquired using a computer-based MA method (implemented in the browser; https://meadows-research.com), in which the participants arranged the images on a screen by mouse drag-and-drop operations. Participants were instructed; "Please arrange these natural scene images according to the similarity of their content" such that the distance between any two images reflected their perceived dissimilarity. This multiple arrangement method uses adaptive selection of image subsets during measurement to estimate the similarity perceived by each individual. During the MA method, a partial representational dissimilarity matrix (RDM; (Kriegeskorte et al., 2008)) is obtained after each trial, and the multiple partial matrices are combined using an iterative weighted averaging (IWA) procedure to provide a single RDM describing the perceived similarity between all pairwise combinations of stimuli for each subject (Kriegeskorte & Mur, 2012). This IWA generated RDM will serve as a reference point to assess the quality of RDMs estimated from our sparse embeddings.

Generating triplets from MA data

For a given trial a triplet of stimuli is selected. The euclidean distance between each pairwise combination of stimuli in this triplet is calculated. The pair with the smallest distance between them is labelled as the most similar pair of items in the triplet. This is repeated for all possible triplet combinations of the trial data, and repeated for every completed trial.

Training Sparse Embeddings

The triplets were used to train a neural network framework initialised with a sparse embedding layer consisting of a 100 x 90 matrix, with each row corresponding to a different item, and each column representing a potential feature dimension (Murphy et al., 2012). For a given stimulus triplet, the pairwise similarity of two items in that triplet is given by the dot product of their corresponding row vectors. The softmax of these dot products is taken as the pair of stimuli most similar in the triplet according to the model. This is repeated for each triplet, and the model predictions are compared with the most similar pairs taken from the MA data. Categorical cross entropy between the two distributions of similarity judgements is taken and backpropagated to update the embedding weights such that they converge to approximate a subject's feature space (Figure 1).

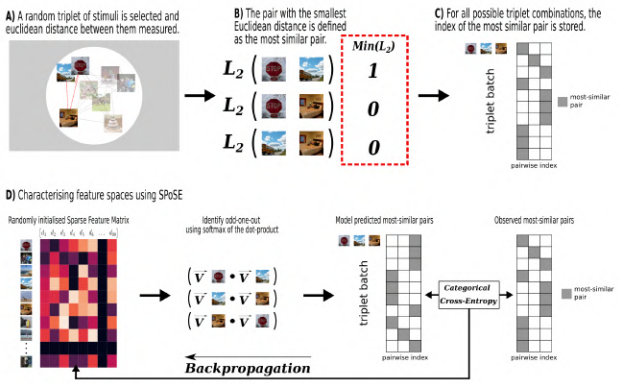

Figure 1: Data from the MA task is used to learn sparse embeddings that predict similarities.

Figure 2: The top 5 stimuli in the 6 most common dimensions across subjects.

Results

Embeddings recover representational geometries

A pairwise dissimilarity value for each item pair was predicted by taking the average probability that a given pair of stimuli is seen as the most similar over all possible triplets in which they occur and subtracting it from 1. RDMs containing model predicted pairwise distances were then compared to IWA RDMs for each participant. The Spearman R values between RDMs were 0.92, 0.90, 0.88, 0.87, 0.87, 0.91, 0.86, 0.85 from subject 1 to subject 8, respectively. On average, this correlation (mean r=0.88) suggests that the model is capable of accurately reproducing the subjects' multidimensional space of similarities.

Inspecting Emergent Feature Dimensions

For each set of trained embeddings, the dimensions seen as most influential were inspected by taking the 5 highest weighted stimuli in each, providing a set of interpretable feature dimensions commonly used in distinguishing between natural scenes. Clear features are obvious across each dimension, relating to animacy, food, transportation, people, indoor and outdoor scenes. The dimensions presented in Figure 2 are those which were most common across all individual subjects.

Emergent Dimensions are Sparse

For each subject; models were re-initialised with a sparse embedding matrix with only 1 potential feature dimension (100 x 1) and trained again. This was repeated multiple times, with an additional dimension added each time such that for each subject, we had 90 sets of embeddings ranging from 100 x 1 to 100 x 90 in shape. Each of these embeddings were used to generate another RDM of model-predicted similarity judgements and compared with that from IWA. This showed that relatively few (6 to 7) dimensions were required to fully explain RDMs for each subject. These findings are consistent with those of Hebart et al. (2020) suggesting that the feature space used in making similarity judgements is sparse, with few dimensions driving these judgements.

Discussion and Future Direction

The aim of this project was to generate a set of interpretable dimensions capable of revealing the key features used when distinguishing between natural scenes. Not only were there clearly observable semantic themes in the highest weighted dimensions, but they were capable of reproducing subject-generated similarity judgements with a high degree of accuracy, suggesting that these dimensions are relevant and utilised by our individual participants. The sparse nature of these dimensions is consistent with the findings of Hebart et al. (2020) indicating that relatively few dimensions are required to fully explain the percieved differences between stimuli. When considering the rich nature of natural scenes, it is worth noting that not all "semantic" features present in each dimension may be immediately observable to the interpreter. The presence of clear, outstanding features such as animacy or setting may occlude our ability to spot more subtle features. Additionally, the use of only 100 images puts a constraint on just how interpretable these dimensions can be, as while the set of stimuli covers a wide range of semantic factors, these factors are not densely populated by stimuli. Interpreting inter-subject idiosyncrasies also suffers as a result of the small stimulus set, as it limits the number of interpretable dimensions in a given subject space to compare with another. Future work should replicate and extend the present findings using a substantially larger set of stimuli.

Acknowledgments

This work was supported by an ERC-StG grant ERC-StG-795468 and an NSERC Discovery grant RGPIN-2021-03127 (to I.C.), and by an NSF CRCNS grant IIS-1822683 (to K.K.).

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Et al. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological review*, *95*(1), 124.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological review*, *93*(2), 154.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder. *CoRR*, *abs/1803.11175*arXiv 1803.11175. http://arxiv.org/abs/1803.11175

Charest, I., Kievit, R., Schmitz, T., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of the beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

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

Kriegeskorte, N., & Mur, M. (2012). Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in psychology*, *3*, 245.

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

Murphy, B., Talukdar, P., & Mitchell, T. (2012). Learning effective and interpretable semantic models using non-negative sparse embedding, In *Proceedings of coling 2012*.