Generalizing Psychological Similarity Spaces to Unseen Stimuli

Combining Multidimensional Scaling with Artificial Neural Networks

Lucas Bechberger and **Kai-Uwe Kühnberger**

1 Introduction

In this chapter, we propose a combination of psychologically derived similarity ratings with modern machine learning techniques in the context of cognitive artificial intelligence. More specifically, we extract a spatial representation of conceptual similarity from psychological data and learn a mapping from visual input onto this spatial representation.

We base our work on the cognitive framework of conceptual spaces (Gärdenfors 2000), which proposes a geometric representation of conceptual structures: Instances are represented as points and concepts are represented as regions in psychological similarity spaces. Based on this representation, one can explain a range of cognitive phenomena from one-shot learning to concept combination. Conceptual spaces can be interpreted as a spatial variant of the influential prototype theory of concepts (Rosch et al. 1976) by identifying the prototype of a given category with the centroid of the respective convex region. Moreover, conceptual spaces can be related to the feature spaces typically used in machine learning (Mitchell 1997), where individual observations are also represented as sets of feature values and where the task is to identify regions which correspond to pre-defined categories.

As Gärdenfors (2018) has argued, the framework of conceptual spaces splits the overall problem of concept learning into two sub-problems: On the one hand, the space itself with its distance relation and its underlying dimensions needs to be

---

The content of this chapter is an updated, corrected, and significantly extended version of research reported in Bechberger and Kypridemou (2018).

L. Bechberger (✉) · K.-U. Kühnberger
Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany
e-mail: lucas.bechberger@uni-osnabrueck.de

K.-U. Kühnberger
e-mail: kai-uwe.kuehnberger@uni-osnabrueck.de

learned. On the other hand, one needs to identify meaningful regions within this similarity space. The latter problem can be easily solved by simple learning mechanisms such as taking the centroid of a given set of category members (Gärdenfors 2000). The problem of obtaining the similarity spaces themselves is however much harder. While in humans, the dimensions of these spaces may be partially innate or learned based on perceptual invariants (Gärdenfors 2018), it is difficult to mimic such processes in artificial systems.

When using conceptual spaces as a modeling tool, one can distinguish three ways of obtaining the underlying dimensions: If the domain of interest is well understood, one can manually define the dimensions and thus the overall similarity space. This can for instance be done for the domain of colors, for which a variety of similarity spaces exists. A second approach is based on machine learning algorithms for dimensionality reduction. For instance, unsupervised artificial neural networks (ANNs) such as autoencoders or self-organizing maps can be used to find a compressed representation for a given set of input stimuli. This task is typically solved by optimizing a mathematical error function which may be not satisfactory from a psychological point of view.

A third way of obtaining the dimensions of a conceptual space is based on dissimilarity ratings obtained from human subjects. One first elicits dissimilarity ratings for pairs of stimuli in a psychological study. The technique of "multidimensional scaling" (MDS) takes as an input these pair-wise dissimilarities as well as the desired number t of dimensions. It then represents each stimulus as a point in a t-dimensional space in such a way that the distances between points in this space reflect the dissimilarities of their corresponding stimuli. *Nonmetric* MDS assumes that the dissimilarities are only ordinally scaled and limits itself to representing the ordering of distances correctly. *Metric* MDS on the other hand assumes an interval or ratio scale and also tries to represent the numerical values of the dissimilarities as closely as possible. We introduce multidimensional scaling in more detail in Sect. 2. Moreover, we present a study investigating the differences between similarity spaces produced by metric and nonmetric MDS in Sect. 3.

One limitation of the MDS approach is that it is unable to generalize to unseen inputs: If a new stimulus arrives, it is impossible to directly map it onto a point in the similarity space without eliciting dissimilarities to already known stimuli. In Sect. 4, we propose to use ANNs in order to learn a mapping from raw stimuli to similarity spaces obtained via MDS. This hybrid approach combines the psychological grounding of MDS with the generalization capability of ANNs.

In order to support our proposal, we present the results of a first feasibility study in Sect. 5: Here, we use the activations of a pre-trained convolutional network as features for a simple regression into the similarity spaces from Sect. 3.

Finally, Sect. 6 summarizes the results obtained in this paper and gives an outlook on future work. Code for reproducing both of our studies can be found online at https://github.com/lbechberger/LearningPsychologicalSpaces/ (Bechberger 2020).

Our overall contribution can be seen as providing artificial systems with a way to map raw perceptions onto psychological similarity spaces. These similarity spaces can then be used in order to learn conceptual regions and to reason with them. Our research has strong relations to two other chapters in this edited volume.

The conceptual spaces framework itself can be considered as a specific instance of the approach labeled as "cognitive and distributional semantics" in the contribution by Färber, Svetashova, and Harth (Chap. 3). Our hybrid proposal from Sect. 4 exemplifies the procedure of obtaining such a cognitive representation which is both psychologically grounded and applicable to novel stimuli. Especially the latter property of our hybrid proposal is crucial for applications in technical systems such as the Internet of Things (IoT) considered by Färber, Svetashova, and Harth.

Also the attribute spaces used by Gust and Umbach (Chap. 4) are closely related to the similarity spaces considered in our contribution. While our work focuses on grounding such a similarity space in perception, Gust and Umbach analyze how natural language similarity expressions can be linked to spatial models. The contribution by Gust and Umbach can thus be seen as a complement to our work, considering a higher level of abstraction.

2 Multidimensional Scaling

In this section, we provide a brief introduction to multidimensional scaling. We first give an overview of the elicitation methods for similarity ratings in Sect. 2.1, before explaining the basics of MDS algorithms in Sect. 2.2. The interested reader is referred to Borg and Groenen (2005) for a more detailed introduction to MDS.

2.1 Obtaining Dissimilarity Ratings

In order to collect similarity ratings from human participants, several different techniques can be used (Goldstone 1994; Hout et al. 2013; Wickelmaier 2003). They are typically grouped into *direct* and *indirect* methods: In direct methods, participants are fully aware that they rate, sort, or classify different stimuli according to their pairwise dissimilarities. Indirect methods on the other hand are based on secondary empirical measurements such as confusion probabilities or reaction times.

One of the classical direct techniques is based on explicit ratings for pairwise comparisons. In this approach, all possible pairs from a set of stimuli are presented to participants (one pair at a time), and participants rate the dissimilarity of each pair on a continuous or categorical scale. Another direct technique is based on sorting tasks. For instance, participants might be asked to group a given set of stimuli into piles of similar items. In this case, similarity is binary—either two items are sorted into the same pile or not.

Perceptual confusion tasks can be used as an indirect technique for obtaining similarity ratings. For example, participants can be asked to report as fast as possible whether two displayed items are the same or different. In this case, confusion probabilities and reaction times are measured in order to infer the underlying similarity relation.

Goldstone ([1994](#)) has argued that the classical approaches for collecting similarity data are limited in various ways. Their biggest shortcoming is that explicitly testing all $\frac{N \cdot (N-1)}{2}$ stimulus pairs is quite time-consuming. An increasing number of stimuli therefore leads to very long experimental sessions which might cause fatigue effects. Moreover, in the course of such long sessions, participants might switch to a different rating strategy after some time, making the collected data less homogeneous.

In order to make the data collection process more time-efficient, Goldstone ([1994](#)) has proposed the "*Spatial Arrangement Method*" (SpAM). In this collection technique, multiple visual stimuli are simultaneously displayed on a computer screen. In the beginning, the arrangement of these stimuli is randomized. Participants are then asked to arrange them via drag and drop in such a way that the distances between the stimuli are proportional to their dissimilarities. Once participants are satisfied with their solution, they can store the arrangement. The dissimilarity of two stimuli is then recorded as their Euclidean distance in pixels. As N items can be displayed at once, each single modification by the user updates N distance values at the same time which makes this procedure quite efficient. Moreover, SpAM quite naturally incorporates geometric constraints: If A and B are placed close together and C is placed far away from A, then it cannot be very close to B.

As the dissimilarity information is recorded in the form of Euclidean distances, one might assume that the dissimilarity ratings obtained through SpAM are ratio scaled. This view is for instance held by Hout et al. ([2014](#)). However, as participants are likely to make only a rough arrangement of the stimuli, this assumption might be too strong in practice. One can argue that it is therefore safer to only assume an ordinal scale. As far as we know, there have been no explicit investigations on this issue. We will provide an analysis of this topic in Sect. [3](#).

2.2   The Algorithms

In this chapter, we follow the mathematical notation by Kruskal ([1964a](#)), who gave the first thorough mathematical treatment of (nonmetric) multidimensional scaling.

One can typically distinguish two types of MDS algorithms (Wickelmaier [2003](#)), namely metric and nonmetric MDS. Metric MDS assumes that the dissimilarities are interval or ratio scaled, while nonmetric MDS only assumes an ordinal scale.

Both variants of MDS can be formulated as an optimization problem involving the pairwise dissimilarities δ_{ij} between stimuli and the Euclidean distances d_{ij} of their corresponding points in the t-dimensional similarity space. More specifically, MDS involves minimizing the so-called "stress" which measures to which extent the spatial representation violates the information from the dissimilarity matrix:

$$\text{stress} = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i<j} \left(d_{ij} \right)^2}}$$

The denominator in this equation serves as a normalization factor in order to make stress invariant to the scale of the similarity space.

In metric MDS, we use $\hat{d}_{ij} = a \cdot \delta_{ij} + b$ to compute stress. This means that we look for a configuration of points in the similarity space whose distances are a linear transformation of the dissimilarities.

In nonmetric MDS, on the other hand, the $\hat{d}_{ij}$ are not obtained by a *linear* but by a *monotone* transformation of the dissimilarities: Let us order the dissimilarities of the stimuli ascendingly: $\delta_{i_1 j_1} < \delta_{i_2 j_2} < \delta_{i_3 j_3} < \ldots$. The $\hat{d}_{ij}$ are then obtained by defining an analogous ascending order, where the difference between the disparities $\hat{d}_{ij}$ and the distances d_{ij} is as small as possible: $\hat{d}_{i_1 j_1} < \hat{d}_{i_2 j_2} < \hat{d}_{i_3 j_3} < \ldots$. Nonmetric MDS therefore only tries to reflect the *ordering* of the dissimilarities in the distances while metric MDS also tries to take into account their differences and ratios.

There are different approaches towards optimizing the stress function, resulting in different MDS algorithms. Kruskal's original nonmetric MDS algorithm (Kruskal 1964b) is based on gradient descent: In an iterative procedure, the derivative of the stress function with respect to the coordinates of the individual points is computed and then used to make a small adjustment to these coordinates. Once the derivative approaches zero, a minimum of the stress function has been found.

A more recent MDS algorithm by de Leeuw (1977) is called SMACOF (an acronym of "**S**caling by **Ma**jorizing a **Co**mplicated **F**unction"). De Leeuw pointed out that Kruskal's gradient descent method has two major shortcomings: Firstly, if the points for two stimuli coincide (i.e., $x_i = x_j$), then the distance function of these two points is not differentiable. Secondly, Kruskal was not able to give a proof of convergence for his algorithm. In order to overcome these limitations, De Leeuw showed that minimizing the stress function is equivalent to maximizing another function λ which depends on the distances and dissimilarities. This function can be easily maximized by using iterative function majorization. Moreover, one can prove that this iterative procedure converges. SMACOF is computationally efficient and guarantees a monotone convergence of stress (Borg and Groenen 2005, Chap. 8).

Picking the right number of dimensions t for the similarity space is not trivial. Kruskal (1964a) proposes two approaches to address this problem.

On the one hand, one can create a so-called "Scree" plot that shows the final stress value for different values of t. If one can identify an "elbow" in this diagram (i.e., a point after which the stress decreases much slower than before), this can point towards a useful value of t.

On the other hand, one can take a look at the interpretability of the generated configurations. If the optimal configuration in a t-dimensional space has a sufficient degree of interpretability and if the optimal configuration in a $t + 1$ dimensional space does not add more structure, then a t-dimensional space might be sufficient.

Fig. 1 Eight example stimuli from the NOUN data set (Horst and Hout 2016)

3 Extracting Similarity Spaces from the NOUN Data Set

It is debatable whether metric or nonmetric MDS should be used with data collected through SpAM. Nonmetric MDS makes less assumptions about the underlying measurement scale and therefore seems to be the "safer" choice. If the dissimilarities are however ratio scaled, then metric MDS might be able to harness these pieces of information from the distance matrix as additional constraints. This might then result in a semantic space of higher quality.

In our study, we compare metric to nonmetric MDS on a data set obtained through SpAM. If the dissimilarities obtained through SpAM are not ratio scaled, then the main assumption of metric MDS is violated. We would then expect that nonmetric MDS yields better solutions than metric MDS. If the dissimilarities obtained through SpAM are however ratio scaled and if the differences and ratios of dissimilarities do contain considerable amounts of additional information, then metric MDS should have a clear advantage over nonmetric MDS.

For our study, we used existing dissimilarity ratings reported for the Novel Object and Unusual Name (NOUN) data set (Horst and Hout 2016), a set of 64 images of three-dimensional objects that are designed to be novel but also look naturalistic. Figure 1 shows some example stimuli from this data set.

3.1 Evaluation Metrics

We used the `stress0` function from R's `smacof` package to compute both metric and nonmetric stress. We expect stress to decrease as the number of dimensions increases. If the data obtained through SpAM is ratio scaled, then we would expect that metric MDS achieves better values on metric stress (and potentially on nonmetric stress as well) than nonmetric MDS. If the SpAM dissimilarities are not ratio scaled, then metric MDS should not have any advantage over nonmetric MDS.

Another possible way of judging the quality of an MDS solution is to look for interpretable directions in the resulting space. However, Horst and Hout (2016) have argued that for the novel stimuli in their data set there are no obvious directions that one would expect. Without a list of candidate directions, an efficient and objective evaluation based on interpretable directions is however hard to achieve. We therefore did not pursue this way of evaluating similarity spaces.

As an additional way of evaluation, we measured the correlation between the distances in the MDS space and the dissimilarity scores from the psychological study.

Pearson's *r* (Pearson 1895) measures the linear correlation of two random variables by dividing their covariance by the product of their individual variances. Given two vectors *x* and *y* (each containing *N* samples from the random variables *X* and *Y*, respectively), Pearson's *r* can be estimated as follows, where $\bar{x}$ and $\bar{y}$ are the average values of the two vectors:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Spearman's ρ (Spearman 1904) generalizes Pearson's *r* by allowing also for nonlinear monotone relationships between the two variables. It can be computed by replacing each observation x_i and y_i with its corresponding rank, i.e., its index in a sorted list, and by then computing Pearson's *r* on these ranks. By replacing the actual values with their ranks, the numeric distances between the sample values lose their importance—only the correct ordering of the samples remains important. Like Pearson's *r*, Spearman's ρ is confined to the interval $[-1, 1]$ with positive values indicating a monotonically increasing relationship.

Both MDS variants can be expected to find a configuration such that there is a monotone relationship between the distances in the similarity space and the original dissimilarity matrix. That is, smaller dissimilarities correspond to smaller distances and larger dissimilarities correspond to larger distances. For Spearman's ρ, we therefore do not expect any notable differences between metric and nonmetric MDS. For metric MDS, we also expect a *linear* relationship between dissimilarities and distances. Therefore, if the dissimilarities obtained by SpAM are ratio scaled, then metric MDS should give better results with respect to Pearson's *r* than nonmetric MDS.

A final way for evaluating the similarity spaces obtained by MDS is visual inspection: If a visualization of a given similarity space shows meaningful structures and clusters, this indicates a high quality of the semantic space. We limit our visual inspection to two-dimensional spaces.

3.2 Methods

In order to investigate the differences between metric and nonmetric MDS in the context of SpAM, we used the SMACOF algorithm in its original implementation in R's `smacof` library.[1] SMACOF can be used in both a *metric* and a *nonmetric* variant. The underlying algorithm stays the same, only the definition of stress and

---

[1] See https://cran.r-project.org/web/packages/smacof/smacof.pdf.

thus the optimization target differs. Both variants were explored in our study. We used 256 random starts with the maximum number of iterations per random start set to 1000. The overall best result over these 256 random starts was kept as final result.

For each of the two MDS variants, we constructed MDS spaces of different dimensionality (ranging from one to ten dimensions). For each of these resulting similarity spaces, we computed both its metric and its nonmetric stress.

In order to analyze how much information about the dissimilarities can be readily extracted from the images of the stimuli, we also introduced two baselines.

For our first baseline, we used the similarity of downscaled images: For each original image (with both a width and height of 300 pixels), we created lower-resolution variants by aggregating all the pixels in a $k \times k$ block into a single pixel (with $k \in [2, 300]$). We compared different aggregation functions, namely, minimum, mean, median, and maximum. The pixels of the resulting downscaled image were then interpreted as a point in a $\lceil \frac{300}{k} \rceil \times \lceil \frac{300}{k} \rceil$ dimensional space.

For our second baseline, we extracted the activation vectors from the second-to-last layer of the pre-trained Inception-v3 network (Szegedy et al. 2016) for each of the images from the NOUN data set. Each stimulus was thus represented by its corresponding activation pattern. While the downscaled images represent surface level information, the activation patterns of the neural network can be seen as more abstract representation of the image.

For each of the three representation variants (downscaled images, ANN activations, and points in an MDS-based similarity space), we computed three types of distances between all pairs of stimuli: The Euclidean distance d_E, the Manhattan distance d_M, and the negated inner product d_{IP}. We only report results for the best choice of the distance function. For each distance function, we used two variants: One where all dimensions are weighted equally and another one where optimal weights for the individual dimensions were estimated based on a non-negative least squares regression in a five-fold cross validation (cf. Peterson et al. (2018) who followed a similar procedure). For each of the resulting distance matrices, we compute the two correlation coefficients with respect to the target dissimilarity ratings. We consider only matrix entries above the diagonal because the matrices are symmetric and all entries on the diagonal are guaranteed to be zero. Our overall workflow is illustrated in Fig. 2.

3.3 Results

Figure 3a shows the Scree plots of the two MDS variants for both metric and nonmetric stress. As one would expect, stress decreases with an increasing number of dimensions: More dimensions help to represent the dissimilarity ratings more accurately. Metric and nonmetric SMACOF yield almost identical performance with respect to both metric and nonmetric stress. This suggests that interpreting the SpAM dissimilarity ratings as ratio scaled is neither helpful nor harmful.
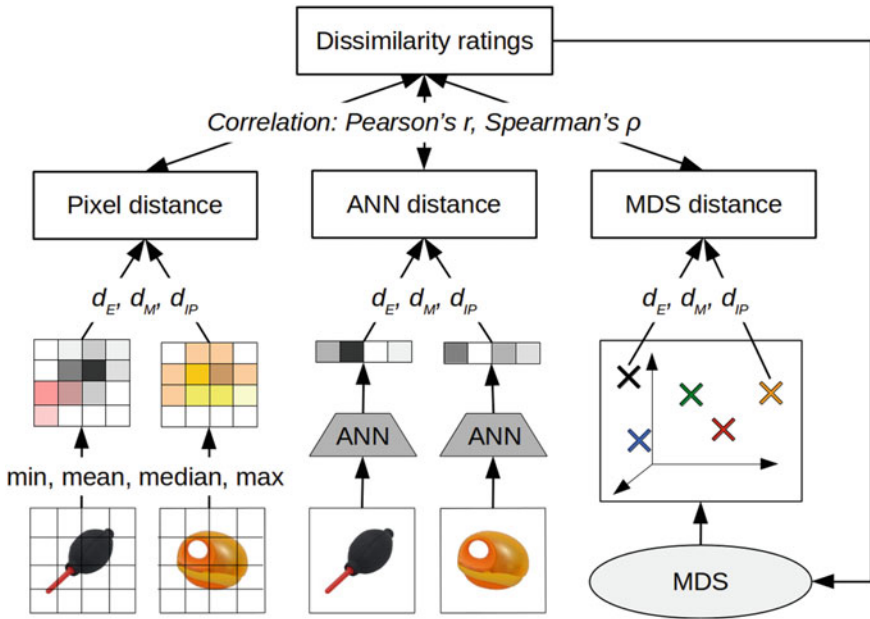
Fig. 2 Illustration of our analysis setup. We measure the correlation between the dissimilarity ratings and distances from three different sources, namely the pixels of downscaled images (left), activations of an artificial neural network (middle), and similarity spaces obtained by MDS (right)

Figure 3b shows some line diagrams illustrating the results of the correlation analysis for the MDS-based similarity spaces. For both the pixel baseline and the ANN baseline, the usage of optimized weights considerably improved performance. As we can see, both of these baselines yield considerably higher correlations than one would expect for randomly generated configurations of points. Moreover, the ANN baseline outperforms the pixel baseline with respect to both evaluation metrics, indicating that raw pixel information is less useful in our scenario than the more high-level features extracted by the ANN. For the pixel baseline, we observed that the minimum aggregator yielded the best results.

We also observe in Fig. 3b that the MDS solutions provide us with a better reflection of the dissimilarity ratings than both pixel-based and ANN-based distances if the similarity space has at least two dimensions. This is not surprising since the MDS solutions are directly based on the dissimilarity ratings, whereas both baselines do not have access to the dissimilarity information. It therefore seems like our naive image-based ways of defining dissimilarities are not sufficient.

With respect to the different MDS variants, also the correlation analysis confirms our observations from the Scree plots: Metric and nonmetric SMACOF are almost indistinguishable with nonmetric SMACOF yielding slightly higher correlation values. This supports the view that the assumption of ratio scaled dissimilarity ratings is not beneficial, but also not very harmful on out data set. Moreover, we find the
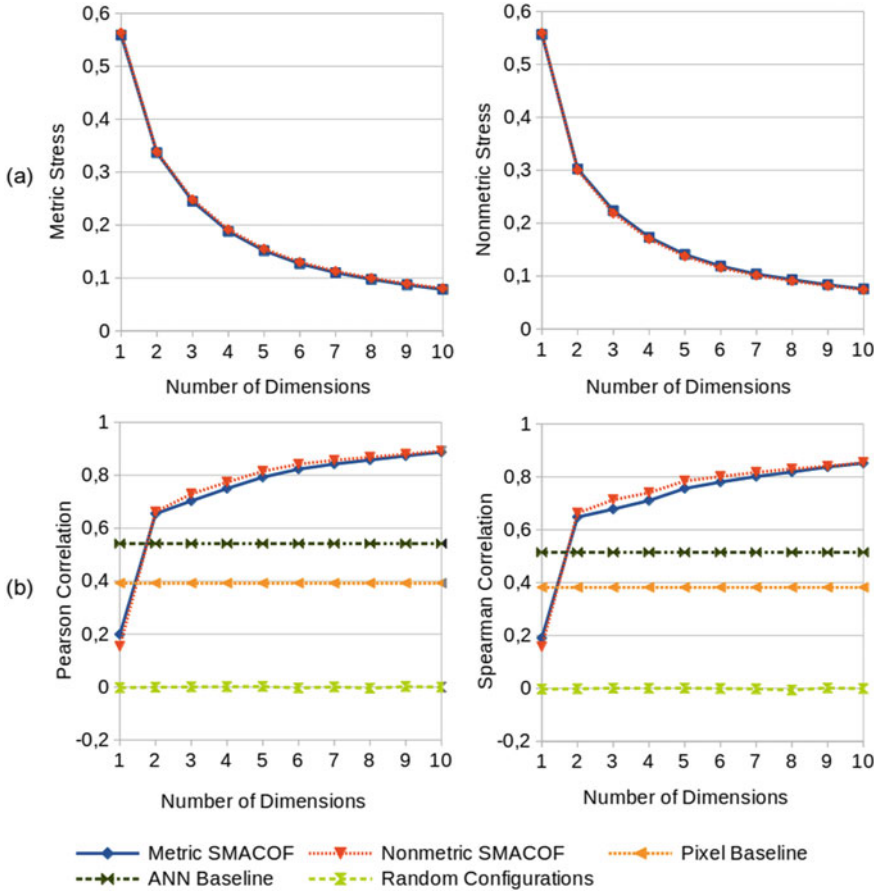
Fig. 3 a Scree plots for both metric and nonmetric stress. **b** Correlation evaluation for the different MDS solutions and the two baselines

tendency of improved performance with an increasing number of dimensions. This again illustrates that MDS is able to fit more information into the space if this space has a larger dimensionality.

Finally, let us look at the two-dimensional spaces generated by the two MDS variants in order to get an intuitive feeling for their semantic structure. Figure 4 shows these spaces along with the local neighborhood of three selected items. These neighborhoods illustrate that in both spaces stimuli are grouped in a meaningful way. From our visual inspection, it seems that both MDS variants result in comparable semantic spaces with a similar structure.

Overall, we did not find any systematic difference between metric and nonmetric MDS on the given data set. It thus seems that the metric assumption is neither beneficial nor harmful when trying to extract a similarity space. On the one hand, we
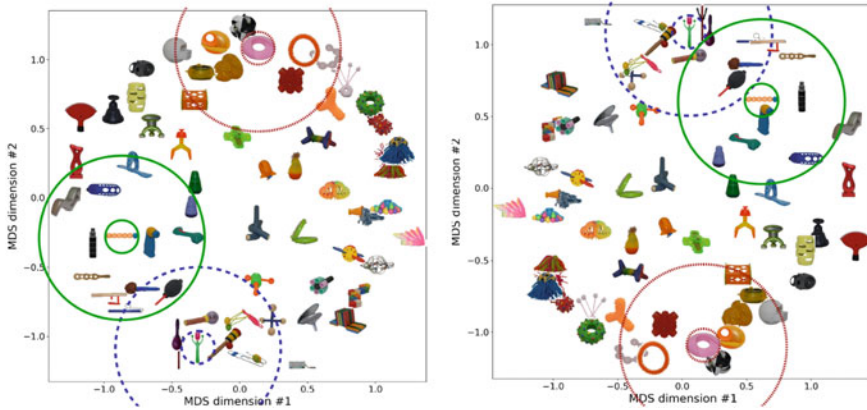
Fig. 4 Illustration of the two-dimensional spaces obtained by metric SMACOF (left) and nonmetric SMACOF (right)

cannot conclude that the dissimilarities obtained through SpAM are *not* ratio scaled. On the other hand, the additional information conveyed by differences and ratios of dissimilarities does not seem to improve the overall results. We therefore advocate the usage of nonmetric MDS due to the smaller amount of assumptions made about the dissimilarity ratings.

4 A Hybrid Approach

Multidimensional scaling (MDS) is directly based on human similarity ratings and leads therefore to conceptual spaces which can be considered psychologically valid. The prohibitively large effort required to elicit such similarity ratings on a large scale however confines this approach to a small set of fixed stimuli. In Sect. 4.1, we propose to use machine learning methods in order to generalize the similarity spaces obtained by MDS to unseen stimuli. More specifically, we propose to use MDS on human similarity ratings to "initialize" the similarity space and artificial neural networks (ANNs) to learn a mapping from stimuli into this similarity space. We afterwards relate our proposal to two other recent studies in this area in Sect. 4.2.

4.1 Our Proposal

In order to obtain a solution having both the psychological validity of MDS spaces and the possibility to generalize to unseen inputs as typically observed for neural networks, we propose the following hybrid approach, which is illustrated in Fig. 5.
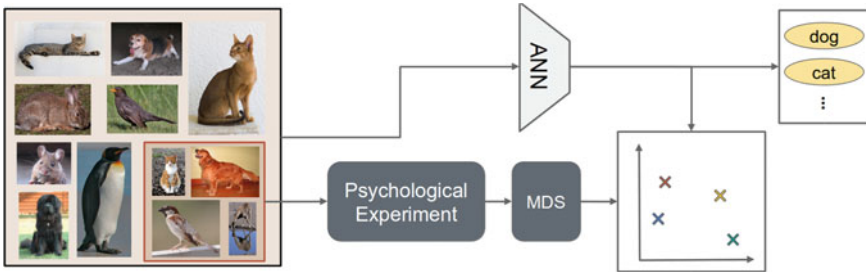
Fig. 5 Illustration of the proposed hybrid procedure: a subset of data is used to construct a conceptual space via MDS. A neural network is then trained to map images into this similarity space, aided by a secondary task (e.g., classification)

After having determined the domain of interest (e.g., the domain of animals), one first needs to acquire a data set of stimuli from this domain. This data set should cover a wide variety of stimuli and it should be large enough for applying machine learning algorithms. Using the whole data set with potentially thousands of stimuli in a psychological experiment is however unfeasible in practice. Therefore, a relatively small, but still sufficiently representative subset of these stimuli needs to be selected for the elicitation of human dissimilarity ratings. This subset of stimuli is then used in a psychological experiment where dissimilarity judgments by humans are obtained, using one of the techniques described in Sect. 2.1.

In the next step, one can apply MDS to these dissimilarity ratings in order to extract a spatial representation of the underlying domain. As stated in Sect. 2.2, one needs to manually select the desired number of dimensions—either based on prior knowledge or by manually optimizing the trade-off between high representational accuracy and a low number of dimensions. The resulting similarity space should ideally be analyzed for meaningful structures and a high correlation of inter-point distances to the original dissimilarity ratings.

Once this mapping from stimuli (e.g., images of animals) to points in a similarity space has been established, we can use it in order to derive a ground truth for a machine learning problem: We can simply treat the stimulus-point mappings as labeled training instances where the stimulus is identified with the input vector and the point in the similarity space is used as its label. We can therefore set up a regression task from the stimulus space to the similarity space.

Artificial neural networks (ANNs) have been shown to be powerful regressors that are capable of discovering highly non-linear relationships between raw low-level stimuli (such as images) and desired output variables. They are therefore a natural choice for this task. ANNs are however a very data-hungry machine learning method — they need large amounts of training examples and many training iterations in order to achieve good performance. On the other hand, the available number of stimulus-point pairs in our proposed procedure is quite low for a machine learning problem — as argued before, we can only look at a small number of stimuli in a psychological experiment.

We propose to resolve this dilemma not only through data augmentation, but also by introducing an additional training objective (e.g., correctly classifying the given images into their respective classes such as CAT and DOG). This additional training objective can also be optimized on all the remaining stimuli from the data set that have not been used in the psychological experiment. Using a secondary task with additional training data constrains the network's weights and can be seen as a form of regularization: These additional constraints are expected to counteract overfitting tendencies, i.e., tendencies to memorize all given mapping examples without being able to generalize.

Figure 5 illustrates the secondary task of predicting the correct classes. This approach is only applicable if the data set contains class labels. If the network is forced to learn a classification task, then it will likely develop an internal representation where all members of the same class are represented in a similar way. The network then "only" needs to learn a mapping from this internal representation (which presumably already encodes at least some aspects of a similarity relation between stimuli) into the target similarity space.

Another secondary task consists in reconstructing the original images from a low-dimensional internal representation, using the structure of an autoencoder. As the computation of the reconstruction error does not require class labels, this is applicable also to unlabeled data sets, which are in general larger and easier to obtain than labeled data sets. The network needs to accurately reconstruct the given stimuli while using only information from a small bottleneck layer. The small size of the bottleneck layer creates an incentive to encode similar input stimuli in similar ways such that the corresponding reconstructions are also similar to each other. Again, this similarity relation learned from the overall data set might be useful for learning the mapping into the similarity space. The autoencoder structure has the additional advantage that one can use the decoder network to generate an image based on a point in the conceptual space. This can be a useful tool for visualization and further analysis.

One should be aware that there is a difference between perceptual and conceptual similarity: Perceptual similarity focuses on the similarity of the raw stimuli, e.g., with respect to their shape, size, and color. Conceptual similarity on the other hand takes place on a more abstract level and involves conceptual information such as the typical usage of an object or typical locations where a given object might be found. For instance, a violin and a piano are perceptually not very similar as they have different sizes and shapes. Conceptually, they might be however quite similar as they are both musical instruments that can be found in an orchestra.

While class labels can be assigned on both the perceptual (ROUND vs. ELONGATED) and the conceptual level (MUSICAL INSTRUMENT vs. FRUIT), the reconstruction objective always operates on the perceptual level. If the similarity data collected in the psychological experiment is of perceptual nature, then both secondary tasks seem promising. If we however target conceptual similarity, then the classification objective seems to be the preferable choice.

4.2 Related Work

Peterson et al. (2017, 2018) have investigated whether the activation vectors of a neural network can be used to predict human similarity ratings. They argue that this can enable researchers to validate psychological theories on large data sets of real world images.

In their study, they used six data sets containing 120 images (each 300 by 300 pixels) of one visual domain (namely, animals, automobiles, fruits, furniture, vegetables, and "various"). Peterson et al. conducted a psychological study which elicited pairwise similarity ratings for all pairs of images using a Likert scale. When applying multidimensional scaling to the resulting dissimilarity matrix, they were able to identify clear clusters in the resulting space (e.g., all birds being located in a similar region of the animal space). Moreover, when applying a hierarchical clustering algorithm on the collected similarity data, a meaningful dendrogram emerged.

In order to extract similarity ratings from five different neural networks, they computed for each image the activation in the second-to-last layer of the network. Then for each pair of images, they defined their similarity as the inner product $(u^T v = \sum_{i=1}^{n} u_i v_i)$ of these activation vectors. When applying MDS to the resulting dissimilarity matrix, no meaningful clusters were observed. Also a hierarchical clustering did not result in a meaningful dendrogram. When considering the correlation between the dissimilarity ratings obtained from the neural networks and the human dissimilarity matrix, they were able to achieve values of R^2 between 0.19 and 0.58 (depending on the visual domain).

Peterson et al. found that their results considerably improved when using a weighted version of the inner product $(\sum_{i=1}^{n} w_i u_i v_i)$: Both the similarity space obtained by MDS and the dendrogram obtained by hierarchical clustering became more human-like. Moreover, the correlation between the predicted similarities and the human similarity ratings increased to values of R^2 between 0.35 and 0.74.

While the approach by Peterson et al. illustrates that there is a connection between the features learned by neural networks and human similarity ratings, it differs from our proposed approach in one important aspect: Their primary goal is to find a way to predict the similarity ratings directly. Our research on the other hand is focused on predicting points in the underlying similarity space.

Sanders and Nosofsky (2018) have used a data set containing 360 pictures of rocks along with an eight-dimensional similarity space for a study which is quite similar in spirit to what we will present in Sect. 5. Their goal was to train an ensemble of convolutional neural networks for predicting the correct coordinates in the similarity space for each rock image from the data set. As the data set is considerably too small for training an ANN from scratch, they used a pre-trained network as a starting point. They removed the topmost layers and replaced them by untrained, fully connected layers with an output of eight linear units, one per dimension of the similarity space. In order to increase the size of their data set, they applied data augmentation methods by flipping, rotating, cropping, stretching, and shrinking the original images.

Their results on the test set showed a value of R^2 of 0.808, which means that over 80% of the variance was accounted for by the neural network. Moreover, an exemplar model on the space learned by the convolutional neural network was able to explain 98.9% of the variance seen in human categorization performance.

The work by Sanders and Nosofsky is quite similar in spirit to our own approach: Like we, they train a neural network to learn the mapping between images and a similarity space extracted from human similarity ratings. They do so by resorting to a pre-trained neural network and by using data augmentation techniques. While they use a data set of 360 images, we are limited to an even smaller data set containing only 64 images. This makes the machine learning problem even more challenging. Moreover, the data set used by Sanders and Nosofky is based on real objects, whereas our study investigates a data set of novel and unknown objects. Finally, while they confine themselves to a single target similarity space for their regression task, we investigate the influence of the target space on the overall results.

5   Machine Learning Experiments

In order to validate whether our proposed approach is worth pursuing, we conducted a feasibility study based on the similarity spaces obtained for the NOUN data set in Sect. 3. Instead of training a neural network from scratch, we limit ourselves to a simple regression on top of a pre-trained image classification network. With the three experiments in our study, we address the following three research questions, respectively:

1. Can we learn a useful mapping from colored images into a low-dimensional psychological similarity space from a small data set of novel objects for which no background knowledge is available?
   Our prediction: The learned mapping is able to clearly beat a simple baseline. However, it does not reach the level of generalization observed in the study of Sanders and Nosofsky (2018) *due to the smaller amount of data available.*
2. How does the MDS algorithm being used to construct the target similarity space influence the results?
   Our prediction: There is are no considerable differences between metric and nonmetric MDS.
3. How does the size of the target similarity space (i.e., the number of dimensions) influence the machine learning results?
   Our prediction: Very small target spaces are not able to reflect the similarity ratings very well and do not contain much meaningful structure. Very large target spaces on the other hand increase the number of parameters in the model which makes overfitting more likely. By this reasoning, medium-sized target spaces should provide a good trade-off and therefore the best regression performance.

5.1 Methods

Please recall from Sect. 3 that the NOUN data base contains only 64 images with an image size of 300 by 300 pixels. As this number of training examples is too low for applying machine learning techniques, we augmented the data set by applying random crops, a Gaussian blur, additive Gaussian noise, affine transformations (i.e., rotations, shears, translations, and scaling), and by manipulating the image's contrast and brightness. These augmentation steps were executed in random order and with randomized parameter settings. For each of the original 64 images, we created 1,000 augmented versions, resulting in a data set of 64,000 images in total. We assigned the target coordinates of the original image to each of the 1,000 augmented versions.

For our regression experiments, we used two different types of feature spaces: The pixels of downscaled images and high-level activation vectors of a pre-trained neural network.

For the ANN-based features, we used the Inception-v3 network (Szegedy et al. 2016). For each of the augmented images, we used the activations of the second-to-last layer as a 2048-dimensional feature vector. Instead of training both the mapping and the classification task simultaneously (as discussed in Sect. 4), we use an already pre-trained network and augment it by an additional output layer.

As a comparison to the ANN-based features, we used an approach similar to the pixel baseline from Sect. 3.2: We downscaled each of the augmented images by dividing it into equal-sized blocks and by computing the minimum (which has shown the best correlation to the dissimilarity ratings in Sect. 3.3) across all values in each of these blocks as one entry of the feature vector. We used block sizes of 12 and 24, resulting in feature vectors of size 1875 and 507, respectively (based on three color channels for downscaled images of size 25×25 and 13×13, respectively). By using these two pixel-based feature spaces, we can analyze differences between low-dimensional and high-dimensional feature spaces. As the high-dimensional feature space is in the same order of magnitude as the ANN-based feature space, we can also make a meaningful comparison between pixel-based features and ANN-based features.

We compare our regression results to the zero baseline which always predicts the origin of the coordinate system. In preliminary experiments, it has shown to be superior to any other simple baselines (such as e.g., drawing from a normal distribution estimated from the training targets). We do not expect this baseline to perform well in our experiments, but it defines a lower performance bound for the regressors.

In our experiments, we limit ourselves to two simple off-the-shelf regressors, namely a linear regression and a lasso regression. Let N be the number of data points, t be the number of target dimensions, $y_d^{(i)}$ the target value of data point i in dimension d, and $f_d^{(i)}$ the prediction of the regressor for data point i in dimension d.

Both of our regressors make use of a simple linear model for each of the dimensions in the target space:

$$f_d = w_0^{(d)} + \sum_{k=1}^{K} w_k^{(d)} x_k$$

Here, K is the number of features and x is the feature vector. In a linear least-squares regression, the weights $w_k^{(d)}$ of this model are estimated by minimizing the mean squared error between the model's predictions and the actual ground truth value:

$$MSE_d = \frac{1}{N} \sum_{i=1}^{N} \left(y_d^{(i)} - f_d^{(i)} \right)^2$$

As the number of features is quite high, even a linear regression needs to estimate a large number of weights. In order to prevent overfitting, we also consider a lasso regression which additionally incorporates the L_1 norm of the weight matrix as regularization term. It minimizes the following objective:

$$\frac{1}{N} \sum_{i=1}^{N} \left(y_d^{(i)} - f_d^{(i)} \right)^2 + \beta \cdot \frac{1}{K} \cdot \sum_{k=1}^{K} w_k^{(d)}$$

The first part of this objective corresponds to the mean squared error of the linear model's predictions, while the second part corresponds to the overall size of the weights. If the constant β is tuned correctly, this can prevent overfitting and thus improve performance on the test set. In our experiments, we investigated the following values:

$$\beta \in \{0.0, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$$

Please note that $\beta = 0$ corresponds to an ordinary linear least-squares regression.

With our experiments, we would also like to investigate whether learning a mapping into a psychological similarity space is easier than learning a mapping into an arbitrary space of the same dimensionality. In addition to the real regression targets (which are the coordinates from the similarity space obtained by MDS), we created another set of regression targets by randomly shuffling the assignment from images to target points. We ensured that all augmented images created from the same original image were still mapped onto the same target point. With this shuffling procedure, we aimed to destroy any semantic structure inherent in the target space. We expect that the regression works better for the original targets than for the shuffled targets.

In order to evaluate both the regressors and the baseline, we used three different evaluation metrics:

- The **mean squared error (MSE)** sums over the average squared difference between the prediction and the ground truth for each output dimension.

$$MSE = \sum_{d=1}^{t} \frac{1}{N} \cdot \sum_{i=1}^{N} \left(y_d^{(i)} - f_d^{(i)} \right)^2$$

- The **mean euclidean distance (MED)** provides us with a way of quantifying the average distance between the prediction and the target in the similarity space.

$$MED = \frac{1}{N} \cdot \sum_{i=1}^{N} \sqrt{\sum_{d=1}^{t} \left(y_d^{(i)} - f_d^{(i)} \right)^2}$$

- The **coefficient of determination** R^2 can be interpreted as the amount of variance in the targets that is explained by the regressor's predictions.

$$R^2 = \frac{1}{t} \cdot \sum_{d=1}^{t} \left(1 - \frac{S_{residual}^{(d)}}{S_{total}^{(d)}} \right) \text{ with } S_{residual}^{(d)} = \sum_{i=1}^{N} \left(y_d^{(i)} - f_d^{(i)} \right)^2$$

$$\text{and } S_{total}^{(d)} = \sum_{i=1}^{N} \left(y_d^{(i)} - \bar{y} \right)^2$$

We evaluated all regressors using an eight-fold cross validation approach, where each fold contains all the augmented images generated from eight of the original images. In each iteration, one of these folds was used as test set, whereas all other folds were used as training set. We aggregated all predictions over these eight iterations (ending up with exactly one prediction per data point) and computed the evaluation metrics on this set of aggregated predictions.

5.2   Experiment 1: Comparing Feature Spaces and Regressors

In our first experiment, we want to test the following hypotheses:

1. The learned mapping is able to clearly beat the baseline. However, it does not reach the level of generalization observed in the study of Sanders and Nosofsky (2018) due to the smaller amount of data available.
2. A regression from the ANN-based features is more successful than a regression from the pixel-based features.
3. As the similarity spaces created by MDS encode semantic similarity by geometric distance, we expect that learning the correct mapping generalizes better to the test set than learning a shuffled mapping.
4. As the feature vectors are quite large, the linear regression has a large number of weights to optimize, inviting overfitting. Regularization through the L_1 loss included in the lasso regressor can help to reduce overfitting.

Table 1 Performance of the different regressors for different feature spaces and correct versus shuffled targets on the four-dimensional space by Horst and Hout (2016). The best results for each combination of column and regressor are highlighted in boldface

Regressor	Feature space	Targets	Test set performance			Degree of overfitting			β
			MSE	MED	R^2	MSE	MED	R^2	
Baseline	Any	Any	1.0000	0.9962	0.0000	1.0000	1.0000	1.0000	–
Linear	ANN (2048)	Correct	**0.6076**	**0.7498**	**0.3766**	42.7093	6.8876	**2.6171**	–
		Shuffled	1.1440	1.0505	−0.1390	56.5298	8.1034	−7.0475	–
	Pixel (1875)	Correct	1.3172	1.0845	−0.3251	2.6191	1.6310	−1.5199	–
		Shuffled	1.6075	1.2099	−0.5955	2.6629	1.6522	−0.6700	–
	Pixel (507)	Correct	1.2073	1.0428	−0.2120	**2.3360**	**1.5433**	−2.2664	–
		Shuffled	1.5853	1.2072	−0.5727	2.5049	1.6036	−0.6458	–
Lasso	ANN (2048)	Correct	**0.5740**	**0.7264**	**0.4111**	28.3766	5.6239	2.3727	0.005, 0.01
	Pixel (1875)	Correct	0.9183	0.9391	0.0788	**1.1320**	1.1371	2.3313	0.2, 0.5
	Pixel (507)	Correct	0.8946	0.9292	0.1015	1.1677	**1.1251**	**2.2538**	0.05, 0.1

5. For smaller feature vectors, we expect less overfitting tendencies than for larger feature vectors. Therefore, less regularization should be needed to achieve optimal performance.

Here, we limit ourselves to a single target space, namely the four-dimensional similarity space obtained by Horst and Hout (2016) through metric MDS.

Table 1 shows the results obtained in our experiment, grouped by the regression algorithm, feature space, and target mapping used. We have also reported the observed degree of overfitting. It is calculated by dividing training set performance by test set performance. Perfect generalization would result in a degree of overfitting of one, whereas larger values reflect the factor to which the regression is more successful on the training set than on the test set. Let us for now only consider the linear regression.

We first focus on the results obtained on the ANN-based feature set. As we can see, the linear regression is able to beat the baseline when trained on the correct targets. The overall approach therefore seems to be sound. However, we see strong overfitting tendencies, showing that there is still room for improvement. When trained on the shuffled targets, the linear regression completely fails to generalize to the test set. This shows that the correct mapping (having a semantic meaning) is easier to learn than an unstructured mapping. In other words, the semantic structure of the similarity space makes generalization possible.

Let us now consider the pixel-based feature spaces. For both of these spaces, we observe that linear regression performs worse than the baseline. Moreover, we can see that learning the shuffled mapping results in even poorer performance than

learning the correct mapping. Due to the overall poor performance, we do not observe very strong overfitting tendencies. Finally, when comparing the two pixel-based feature spaces, we observe that the linear regression tends to perform better on the low-dimensional feature space than on the high-dimensional one. However, these performance differences are relatively small.

Overall, ANN-based features seem to be much more useful for our mapping task than the simple pixel-based features, confirming our observations from Sect. 3.

In order to further improve our results, we now varied the regularization factor β of the lasso regressor for all feature spaces.

For the ANN-based feature space, we are able to achieve a slight but consistent improvement by introducing a regularization term: Increasing β causes poorer performance on the training set while yielding improvements on the test set. The best results on the test set are achieved for $\beta \in \{0.005, 0.01\}$. If β however becomes too large, then performance on the test set starts to decrease again — for $\beta = 0.05$ we do not see any improvements over the vanilla linear regression any more. For $\beta \geq 5$, the lasso regression collapses and performs worse than the baseline.

Although we are able to improve our performance slightly, the gap between training set performance and test set performance still remains quite high. It seems that the overfitting problem can be somewhat mitigated but not solved on our data set with the introduction of a simple regularization term.

When comparing our best results to the ones obtained by Sanders and Nosofsky (2018) who achieved values of $R^2 \approx 0.8$, we have to recognize that our approach performs considerably worse with $R^2 \approx 0.4$. However, the much smaller number of data points in our experiment makes our learning problem much harder than theirs. Even though we use data augmentation, the small number of different targets might put a hard limit on the quality of the results obtainable in this setting. Moreover, Sanders and Nosofsky retrained the whole neural network in their experiments, whereas we limit ourselves to the features extracted by the pre-trained network. As we are nevertheless able to clearly beat our baselines, we take these results as supporting the general approach.

For the pixel-based feature spaces, we can also observe positive effects of regularization. For the large space, the best results on the test set are achieved for larger values of $\beta \in \{0.2, 0.5\}$. These results are however only slightly better than baseline performance. For the small pixel-based feature space, the optimal value of β lies in $\{0.05, 0.1\}$, leading again to a test set performance slightly superior to the baseline. In case of the small pixel-based feature space, already values of $\beta \geq 1$ lead to a collapse of the model.

Comparing the regularization results on the three feature spaces, we can conclude that regularization is indeed helpful, but only to a small degree. On the ANN-based feature space, we still observe a large amount of overfitting, and performance on the pixel-based feature spaces is still relatively close to the baseline. Looking at the optimal values of β, it seems like the lower-dimensional pixel-based feature space needs less regularization than its higher-dimensional counterpart. Presumably, this is caused by the smaller possibility for overfitting in the lower-dimensional feature space. Even though the larger pixel-based feature space and the ANN-based feature

space have a similar dimensionality, the pixel-based feature space requires a larger degree of regularization for obtaining optimal performance, indicating that it is more prone to overfitting than the ANN-based feature space.

5.3 Experiment 2: Comparing MDS Algorithms

After having analyzed the soundness of our approach in experiment 1, we compare target spaces of the same dimensionality, but obtained with different MDS algorithms. More specifically, we compare the results from experiment 1 to analogous procedures applied to the ANN-based feature space and the four-dimensional similarity spaces created by both metric and nonmetric SMACOF in Sect. 3. Table 2 shows the results of our second experiment.

In a first step, we can compare the different target spaces by taking a look at the behavior of the zero baseline in each of them. As we can see, the values for MSE and R^2 are identical for all of the different spaces. Only for the MED we can observe some slight variations, which can be explained by the slightly different arrangements of points in the different similarity spaces.

As we can see from Table 2, the results for the linear regression on the different target spaces are comparable. This adds further support to our results from Sect. 3:

Table 2 Comparison of the results obtainable on four-dimensional spaces created by different MDS algorithms. Best results in each column are highlighted for each of the regressors

Regressor	Target space	Test set performance			Amount of overfitting			β
		MSE	MED	R^2	MSE	MED	R^2	
Baseline	Horst and Hout	1.0000	0.9962	0.0000	1.0000	1.0000	1.0000	–
	Metric SMACOF	1.0000	0.9981	0.0000	1.0000	1.0000	1.0000	–
	Nonmetric SMACOF	1.0000	**0.9956**	0.0000	1.0000	1.0000	1.0000	–
Linear	Horst and Hout	**0.6076**	0.7498	**0.3766**	42.7093	6.8876	2.6171	–
	Metric SMACOF	0.6172	0.7560	**0.3766**	**42.2885**	6.8583	**2.6162**	–
	Nonmetric SMACOF	0.6086	**0.7461**	0.3706	42.4380	**6.8305**	2.6585	–
Lasso	Horst and Hout	**0.5740**	**0.7264**	**0.4111**	**28.3766**	**5.6239**	**2.3727**	0.005, 0.01
	Metric SMACOF	0.6052	0.7458	0.3880	35.0367	6.2463	2.5326	0.002
	Nonmetric SMACOF	0.5938	0.7316	0.3853	29.1236	5.6497	2.5413	0.005

Also when considering the usage as target space for machine learning, metric MDS does not seem to have any advantage over nonmetric MDS.

For the lasso regressor, we observed similar effects for all of the target spaces: A certain amount of regularization is helpful to improve test set performance, while too much emphasis on the regularization term causes both training and test set performance to collapse. We still observe a large amount of overfitting even after using regularization. Again, the results are comparable across the different target spaces. However, the optimal performance on the space obtained with metric SMACOF is consistently worse than the results obtained on the other two spaces. As the space by Horst and Hout is however also based on metric MDS, we cannot use this observation as an argument for nonmetric MDS.

5.4 Experiment 3: Comparing Target Spaces of Different Size

In our third and final experiment in this study, we vary the number of dimensions in the target space. More specifically, we consider similarity spaces with one to ten dimensions that have been created by nonmetric SMACOF. Again, we only consider the ANN-based feature space.

Table 3 displays the results obtained in our third experiment and Fig. 6 provides a graphical illustration. When looking at the zero baseline, we observe that the mean Euclidean distance tends to grow with an increasing number of dimensions, with an asymptote of one. This indicates that in higher-dimensional spaces, the points seem to lie closer to the surface of a unit hypersphere around the origin. For both MSE and R^2, we do not observe any differences between the target spaces.

Let us now look at the results of the linear regression. It seems that for all the evaluation metrics, a two-dimensional target space yields the best result. With an increasing number of dimensions in the target space, performance tends to decrease. We can also observe that the amount of overfitting is optimal for a two-dimensional space and tends to increase with an increasing number of dimensions. A notable exception is the one-dimensional space which suffers strongly from overfitting and whose performance with respect to all three evaluation metrics is clearly worse than the baseline.

The optimal performance of a lasso regressor on the different target spaces yields similar results: For all target spaces, a certain amount of regularization can help to improve performance but too much regularization decreases performance. Again, we can only counteract a relatively small amount of the observed overfitting. As we can see in Table 3, again a two-dimensional space yields the best results. With respect to the optimal regularization factor β, we can observe that low-dimensional spaces with up to three dimensions seem to use larger values of β than higher-dimensional spaces with four dimensions and more. This difference in the degree of regularization is also reflected in the different degrees of overfitting observed for these groups of spaces.

Table 3 Performance of the zero baseline, the linear regression, and the lasso regression on target spaces of different dimensionality t derived with nonmetric SMACOF, along with the relative amount of overfitting. Best values for each column are highlighted for each of the regressors

Regressor	t	Test set performance			Amount of overfitting			β
		MSE	MED	R^2	MSE	MED	R^2	
Baseline	1	1.0000	**0.8664**	0.0000	1.0000	1.0000	1.0000	–
	2	1.0000	0.9580	0.0000	1.0000	1.0000	1.0000	–
	3	1.0000	0.9848	0.0000	1.0000	1.0000	1.0000	–
	4	1.0000	0.9956	0.0000	1.0000	1.0000	1.0000	–
	5	1.0000	0.9966	0.0000	1.0000	1.0000	1.0000	–
	6	1.0000	0.9973	0.0000	1.0000	1.0000	1.0000	–
	7	1.0000	0.9978	0.0000	1.0000	1.0000	1.0000	–
	8	1.0000	0.9980	0.0000	1.0000	1.0000	1.0000	–
	9	1.0000	0.9982	0.0000	1.0000	1.0000	1.0000	–
	10	1.0000	0.9984	0.0000	1.0000	1.0000	1.0000	–
Linear	1	1.1499	0.9046	−0.1499	59.0040	8.3419	−6.5413	–
	2	**0.4995**	**0.6370**	**0.5002**	38.9046	**6.5291**	**1.9734**	–
	3	0.5554	0.6979	0.4435	41.4309	6.7360	2.2243	–
	4	0.6086	0.7461	0.3706	42.4380	6.8305	2.6585	–
	5	0.6333	0.7692	0.3595	43.4577	6.9023	2.7405	–
	6	0.6359	0.7734	0.3469	43.4900	6.8770	2.8397	–
	7	0.6675	0.7956	0.3204	44.7364	6.9621	3.0741	–
	8	0.6846	0.8094	0.3033	45.1247	6.9876	3.2459	–
	9	0.6810	0.8078	0.2983	44.8367	6.9591	3.3004	–
	10	0.7107	0.8259	0.2807	46.0530	7.0432	3.5076	–
Lasso	1	0.9912	0.8368	0.0088	**1.3656**	**1.7043**	30.9878	1, 2
	2	**0.4728**	**0.6052**	**0.5271**	19.1298	4.5081	**1.8504**	0.02
	3	0.5322	0.6720	0.4722	19.4148	4.5725	2.0593	0.02
	4	0.5938	0.7316	0.3853	29.1237	5.6497	2.5413	0.005
	5	0.6180	0.7576	0.3755	35.1383	6.2167	2.6160	0.002
	6	0.6274	0.7651	0.3548	35.0732	6.1797	2.7724	0.001, 0.002
	7	0.6589	0.7839	0.3280	39.6352	5.0619	2.9979	0.001, 0.01
	8	0.6752	0.8022	0.3117	39.5496	6.5669	3.1527	0.001
	9	0.6680	0.7980	0.3108	38.8777	6.1359	3.1615	0.001, 0.002
	10	0.6993	0.8166	0.2924	35.3561	5.5563	3.3513	0.002, 0.005
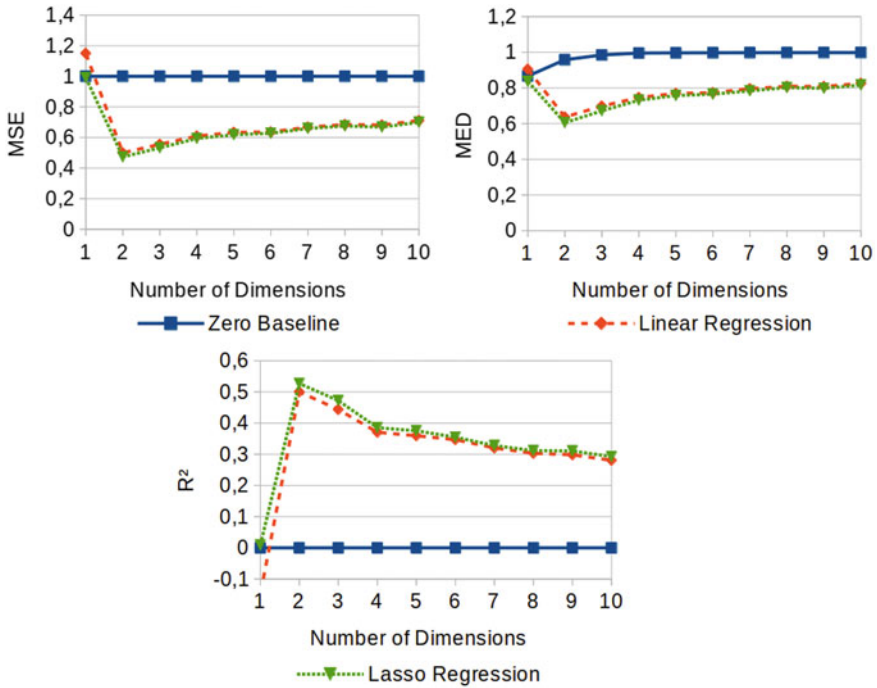
Fig. 6 Visualization of the regression results for MSE, MED, and R^2 as a function of the number of dimensions

Taken together, the results of our third experiment show that a higher-dimensional target space makes the regression problem more difficult, but that a one-dimensional target space does not contain enough semantic structure for a successful mapping. It seems that a two-dimensional space is in our case the optimal trade-off. However, even the performance of the lasso regressor on this space is far from satisfactory, urging for further research.

6   Conclusions

The contributions of this paper are twofold.

In our first study, we investigated whether the dissimilarity ratings obtained through SpAM are ratio scaled by applying both metric MDS (which assumes a ratio scale) and nonmetric MDS (which only assumes an ordinal scale). Both MDS variants produced comparable results—it thus seems that assuming a ratio scale is neither beneficial nor harmful. We therefore recommend to use nonmetric MDS as its underlying assumptions are weaker. Future studies on other data sets obtained through SpAM should seek to confirm or contradict our results.

In our second study, we analyzed whether learning a mapping from raw images to points in a psychological similarity space is possible. Our results showed that using the activations of a pre-trained ANN as features for a regression task seems to work in principle. However, we observed very strong overfitting tendencies in our experiments. Furthermore, the overall performance level we were able to achieve is still far from satisfactory. The results by Sanders and Nosofsky (2018) however show that larger amounts of training data can alleviate these problems. Future work in this area should focus on improvements in performance and robustness of this approach.

As follow-up work, we are currently conducting a study on a data set of shapes, where we plan to apply more sophisticated machine learning methods in order to counteract the observed overfitting tendencies.

References

Bechberger, L. (2020). lbechberger/LearningPsychologicalSpaces v1.3: Study on multidimensional scaling and neural networks on the NOUN dataset. https://doi.org/10.5281/zenodo.4061287.

Bechberger, L., & Kypridemou, E. (2018). Mapping images to psychological similarity spaces using neural networks. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Cognition*.

Borg, I., & Groenen, J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). Springer Series in Statistics. New York: Springer.

de Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In *Recent development in statistics* (pp. 133–146). North Holland Publishing.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press.

Gärdenfors, P. (2018). From sensations to concepts: A proposal for two learning processes. *Review of Philosophy and Psychology*, *10*(3), 441–464.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386.

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409.

Hout, M. C., Goldinger, S. D., & Brady, K. J. (2014). MM-MDS: A multidimensional scaling database with similarity ratings for 240 object categories from the massive memory picture database. *PLOS ONE*, *9*(11), 1–11.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, *142*(1), 256.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347–352), 240–242.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2017). Adapting deep network features to capture psychological representations: An abridged report. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (pp. 4934–4938).

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

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Sanders, C. A., & Nosofsky, R. M. (2018). Using deep-learning representations of complex natural stimuli as input to psychological models of classification. In *Proceedings of the 2018 Conference of the Cognitive Science Society, Madison*.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).

Wickelmaier, F. (2003). An introduction to MDS. *Sound Quality Research Unit, Aalborg University, Denmark*, *46*(5), 1–26.