Constructing a Psychological Feature Space using Deep Neural Networks

Jayesh Prasad Anandan (janandan@iu.edu)[*], Bhakti Patrawala (bhpatra@iu.edu)[*], Afeefa Bano (abano@iu.edu)[*]

Abstract
Building upon prior research efforts centered on constructing a psychological feature space for natural object categories using deep neural networks, this study aims to expand and replicate the experiment by employing alternative neural network architectures. By deviating from the original approach, we aim to investigate the impact of different network structures on the construction of psychologically grounded feature spaces. Through rigorous experimentation and analysis, this project endeavors to assess the efficacy and nuances of varied neural network architectures in capturing and refining representations aligned with psychological dimensions of natural object categories. The outcomes of this study are anticipated to contribute valuable insights into the adaptability and robustness of neural network models in generating comprehensive feature spaces relevant to human perception, thereby advancing the understanding of AI-based object categorization in alignment with psychological attributes.
Project Github Link: `https://github.iu.edu/janandan/DeepLearningSystemsProject`

Keywords
Psychological Feature Space, Natural Object Categories, Neural Network Architectures, Representation Learning, Object Categorization, Psychological Dimensions, Multidimensional Scaling

[*]*Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1. Introduction

This research project delves into the exploration of constructing a psychological feature space for natural object categories through the utilization of deep neural networks. It investigates the impact of varied neural network architectures on the creation of representations aligned with psychological dimensions of objects. By employing alternative architectures, the study aims to replicate and extend prior experiments, seeking to understand the subtleties and efficacy of diverse network structures in capturing nuanced psychological attributes associated with object categories. The research focuses on the intersection of representation learning, psychological dimensions, and AI perception, aiming to advance the understanding of object categorization by uncovering how neural networks adapt and refine feature spaces that resonate with human perception and cognitive processes.

2. Method

2.1 Dataset

The dataset we used in our project is sourced from Nosofsky et al. (*2018c* [1]), which consists of 360 images of rocks. The dataset belongs to three higher level categories of Igneous, Metamorphic and sedimentary. Each of these higher level categories contain ten subtypes and twelve individual tokens within each subtype.
The subtypes are as follows:

Table 1. Types and Subtypes of Rocks Dataset

Igneous	Metamorphic	Sedimentary
Andesite	Amphibolite	Bituminous coal
Basalt	Anthracite	Breccia
Diorite	Gneiss	Chert
Gabbro	Hornfels	Conglomerate
Granite	Marble	Dolomite
Obsidian	Migmatite	Micrite
Pegmatite	Phyllite	Rock gypsum
Peridotite	Quartzite	Rock salt
Pumice	Schist	Sandstone
Rhyolite	Slate	Shale

The dataset also contains feature values for each of 360 images in 8 psychological dimensions, calculated using Multi-Dimensional Scaling (MDS) (*Nosofsky et al. 2018c* [1], *2019a* [2]). Participants judged similarities between pairs of rock images, generating a large dataset used to create a spatial arrangement. This arrangement was refined through statistical procedures and aligned with independent data to enhance interpretability, ultimately determining the final dimensions based on fit and understandability.

2.2 Training and Procedure

2.2.1 Replication

From the experiment conducted by *Sanders, C.A., Nosofsky, R.M. (2020)* [3], the challenge of a small dataset with only 360 rocks was addressed using transfer learning—a technique leveraging a pre-trained ResNet50 (*He et al. 2016* [4]) model initially built on ILSVRC dataset. By modifying the model's output layer and employing data augmentation techniques like flipping, rotating, and cropping, the training set was expanded artificially. The goal was to minimize error by training the modified model to align its output with the multi-dimensional scaling (MDS) coordinates of the rocks, ensuring equal weight across the eight dimensions.

To enhance accuracy, an ensemble of 10 convolutional neural networks (CNNs) was created, recognizing that varying initial parameters can lead to different outcomes. By averaging the outputs of these networks, promising validation results were achieved (MSE = 1.298, R2 = 0.780). However, to gauge unbiased generalization performance, the ensemble's efficacy was tested using a separate dataset—the test set. This evaluation provided a more accurate estimate of the model's true ability to generalize beyond the training data.

The ensemble of CNNs produced an (MSE = 1.355, R2 = 0.767) on the test set, indicating that it captures more than 75% of the variance in both validation and test sets. This suggests a strong initial indication that properly trained deep learning networks have the potential to automatically derive psychological representations from natural stimuli.

To test generalization, a new MDS study involving 120 rocks, similar to the initial 360-rock set, was conducted. This study aimed to observe if the same dimensions emerged and assess whether CNNs could generalize to this independent rock set. Similarity ratings between these new rocks were collected, along with independent ratings for characteristics like color, grain size, roughness, and others. Using these, an 8-dimensional MDS space was created and aligned with dimension ratings to aid interpretation, following a methodology from (*Nosofsky et al. 2018c*[1], *2019b* [5])

2.2.2 Extension To DenseNet

Based on the paper's recommendations for enhancing Deep Neural Network performance to derive MDS values, we endeavored to implement DenseNet (*Huang, G et al. 2017* [6]). DenseNet's advantages, including reduced parameter count and potential efficacy with limited training data, align well with our objective. Its dense connectivity pattern facilitates superior feature propagation and augments information flow, especially crucial for gradient flow and parameter efficiency.

Our approach involved modifying the DenseNet output layer and incorporating data augmentation techniques to expand our training dataset of 180 images. Employing an ensemble method, we constructed 10 models employing Mean Squared Error (MSE) as the Loss function. Aggregating these model outputs allowed us to compute MSE and R2 Scores for both Validation and Test Sets. Remarkably, with minimal hyperparameter tuning and model complexity exploration, our results closely paralleled those in the paper. Leveraging DenseNet for capturing low-level features and fine-tuning through Dense layers, along with Nadam Optimizer and He Normal weight initialization techniques—learned from our Deep Learning Systems course—facilitated faster model convergence.

Our model exhibited promising performance on unseen data, yielding (MSE = 1.315, R2 = 0.768) on the validation set and (MSE = 1.443, R2 = 0.752) on the test set. These outcomes were obtained using 90 images in both the validation and test sets, ensuring robustness beyond the training data.

Building upon the successes observed with DenseNet, our ongoing pursuit for enhanced performance has led us to explore the potential of Vision Transformer (ViT) (*Alexey Dosovitskiy et al. 2020* [7]) models within this project. While our initial endeavors with DenseNet have yielded commendable results, the inherent strengths of ViT models in capturing global dependencies and leveraging self-attention mechanisms prompt us to further investigate their applicability in our context.

2.2.3 Extension To Vision Transformers

In pursuit of this goal, we constructed a vision transformer model for Image Classification using Hugging Face and Keras. The model processes images by converting them into a linearly embedded sequence of image patches, supplemented with a unique token positioned at the sequence's outset, aiding in image classification. The incorporation of positional embeddings further enhances this sequence, enriching the input fed into the model. For our experimentation, we fine-tuned the google/vit-base-patch16-224-in21k (*Bichen Wu et al. 2020* [8])

Vision Transformer, initially pretrained on ImageNet-21k—a colossal dataset comprising 14 million images across 21843 classes—all standardized at a resolution of 224x224.

We harnessed the Transformers feature extractor to process images stored in folders, transforming them into pixel values while applying augmentation techniques to prepare them for input into the transformer architecture. Subsequently, we curated and partitioned our dataset, allocating 240 images for training and 120 for validation. This segmentation enabled us to fine-tune our hyperparameters effectively.

Our process involved loading the pre-existing model by converting our dataset into a TensorFlow Dataset and employing a Data Collator. Augmenting this setup, we extended the Vision Transformer Model. After the ViT model, we introduced an additional layer—our MDS Dimensions Layer—activated by tanh. To classify the images across the three distinct classes within our dataset, we appended an output layer employing softmax activation.

For training, we employed Sparse Categorical Cross Entropy as our Loss function and Adam Optimizer. Training sessions persisted for 20 epochs, utilizing default hyperparameter values akin to those in the pre-trained model. This approach enabled us to iteratively refine our model's performance and attain optimal classifications across the dataset's distinct classes.

Employing the early stopping mechanism to mitigate overfitting, we attained a commendable validation accuracy of 68.33%. However, since this is just the image classification accuracy, we are focused on generating MDS values close to the MDS values generated by Humans. Subsequently, leveraging the learned model weights, we computed the activations within our MDS layer. This enabled us to predict the activations across our entire dataset of 360 images. Furthermore, extending this capability to a separate collection of 120 entirely new images, we utilized the model to forecast the activations within this distinct dataset.

We utilized the Procrustes method to compare the MDS values generated by the Vision Transformer (ViT) with the MDS values obtained from the research paper. This comparison allowed us to compute the Disparity between the two sets of values. For the 360 Rock Images Dataset, the Disparity from ViT-generated MDS values equaled 0.819, whereas for the 120 Images Dataset, it measured 0.803.

Moreover, to comprehensively evaluate the performance of the Vision Transformer across all dimensions, we calculated the Average Pearson Correlation Coefficient. This metric assessed the correlation between the predicted MDS values (from the ViT model) and the actual MDS Values. For the 360 Image Dataset, the Average Correlation Coefficient was determined to be 0.3949. Similarly, for the 120 Image Dataset, the Average Correlation Coefficient stood at 0.4202. These coefficients provide insights into the model's performance in capturing the relationships and patterns across various dimensions within the datasets

2.3 Application

2.3.1 Generalization within original Space

The below plot shows us the correlation between the ensemble's predictions and the actual MDS Values.

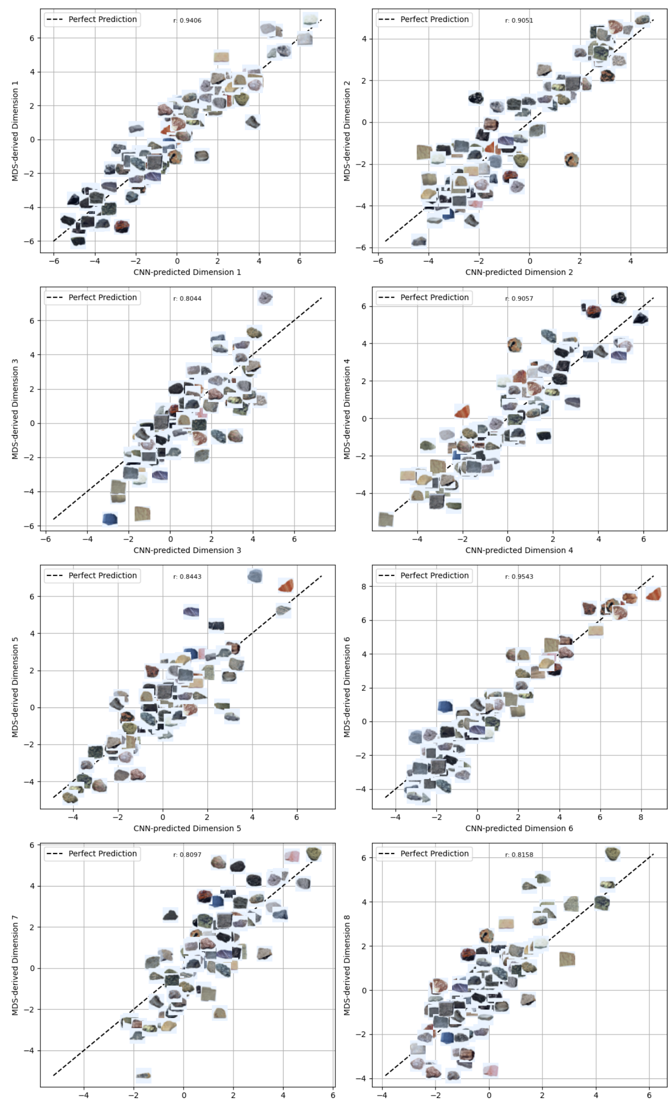

Figure 1. Scatter Plot of the actual MDS Values of Test Set

This analysis reveals a noteworthy correlation between the ensemble's predictions and the actual MDS values, particularly demonstrating high accuracy across most dimensions. The CNNs excel notably in interpreting dimensions related to lightness (Dimension 1) and chromaticity (Dimension 6), aligning with their strength in capturing low-level color information. However, their performance diminishes when confronted with the "shape" dimension (Dimension 7), which lacks a clear interpretation. Nonetheless, the CNNs' ability to produce reasonably accurate predictions in this dimension is intriguing, suggesting underlying meaning that may not be immediately evident to human observers.

A surprising observation lies in the CNNs' relatively poor performance on the roughness dimension (Dimension 3), akin to their performance on the shape dimension, despite the latter having a seemingly clearer interpretation. Upon scrutinizing mispredicted rocks, it becomes apparent that certain rocks positioned in the smooth section of the MDS space possess bumpy or wavy textures, suggesting roughness not fully reflected in their MDS coordinates. This discrepancy hints at potential noise within the derived MDS space, which is expected considering its basis on an incomplete similarity matrix.

2.3.2 Generalization outside the original Space

It's necessary to test models on new data to ensure they generalize well. However, our concern arises from the fact that our test set shares the same MDS space as the training and validation sets. There's uncertainty whether using a new set of rocks from the same categories would yield similar dimensions in the MDS analysis. This uncertainty challenges the CNNs' ability to generalize effectively. Hence, we conducted an MDS study with 120 new rocks from the same categories as the original set to assess if similar dimensions would emerge and to evaluate the CNNs' capability to generalize to this distinct rock collection.

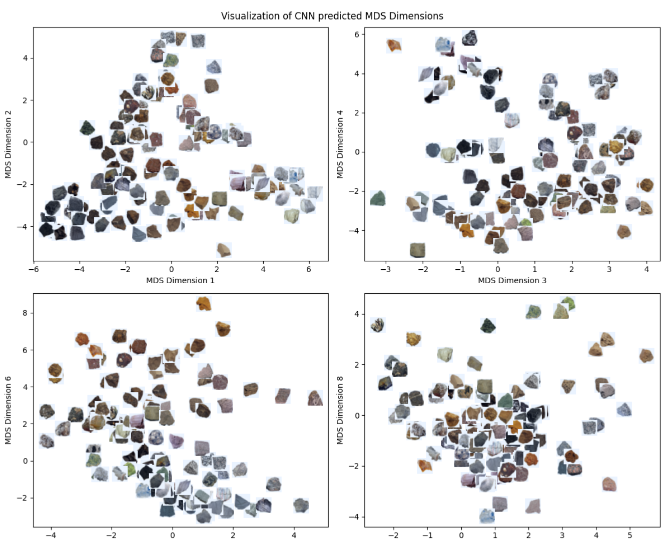

Figure 2. Rotated MDS Space for the 120 Image Set

Figure 2 exhibits the rotated MDS space, while Figure 3 showcases scatterplots depicting the relationship between these MDS dimensions and the 8 predicted dimensions derived from the ensemble of CNNs specifically for the 120 Images Dataset.

The analysis of these figures highlights certain MDS dimensions interpretability in the 120-rock MDS space, akin to the original 360-rock MDS space. Specifically, dimensions 1, 2, 4, and 6 align with lightness/darkness, average grain size, shininess, and chromaticity, respectively. Notably, strong correlations exist between these MDS dimensions, direct dimension ratings, and those predicted by the CNN ensemble, affirming the ensemble's capacity to generalize even beyond the trained MDS space.

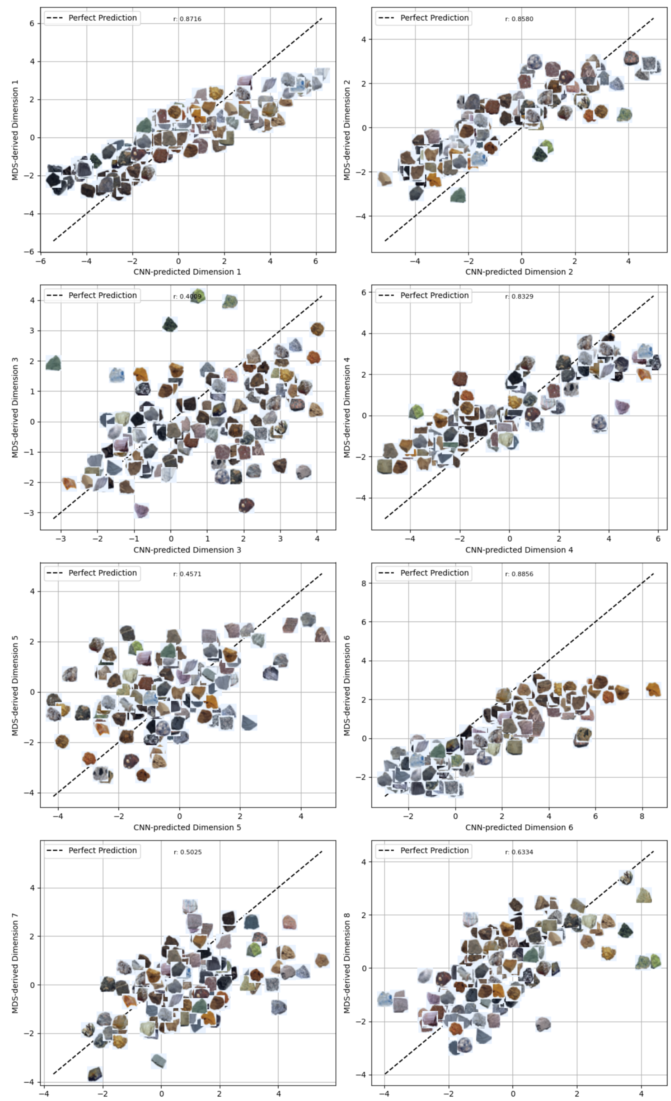

Figure 3. Scatter Plot of the actual MDS Values of 120 Image Set

However, the interpretations of dimensions 3, 5, and 7 in the 120-rock MDS space lack clarity compared to their definition in the 360-rock MDS space. Although some associations between roughness and organization exist, exceptions dilute these connections, resulting in moderate correlations with direct ratings and CNN-predicted dimensions. Despite this, the CNN predictions align reasonably well with the observed trends, indicating a degree of meaningfulness, albeit with differences in the derived MDS spaces across studies.

Remarkably, dimension 7, initially lacking a clear interpretation in the 360-rock MDS space, reemerges in the 120-rock MDS space. The observed shape differences between flat and spherical/cubical rocks influence participants' similarity ratings, showcasing the CNNs' ability to capture such nuances, albeit with modest correlation. This consistent emergence across MDS spaces underscores its psychological significance, warranting further investigation for a comprehensive interpretation.

This analysis underscores the CNNs' capability to generalize MDS dimensions to entirely new rock datasets, demonstrating promising potential for generalization beyond the trained dataset of (*Nosofsky et al. 2018c* [1]).

3. Results

1. ResNet Model as observed in *Sanders, C.A., Nosofsky, R.M. (2020)* [9]

Table 2. Performance of ResNet Model

Dataset	MSE	R2 Score
Validation	1.298	0.780
Test	1.355	0.767

2. DenseNet Model

Table 3. Performance of DenseNet Model

Dataset	MSE	R2 Score
Validation	1.3152	0.7687
Test	1.4433	0.7525

3. Vision Transformer using Pretrained Model

Table 4. Average Correlation Coefficient

Dataset	Average Corr. Coeff.
360 Images	0.39489
120 Images	0.42019

Within the constraints of our project timeline, our focus on enhancing the DenseNet Architecture yielded commendable results, positioning it on par with the ResNet model referenced in the paper we sought to improve upon. This observation sheds light on the potential of the DenseNet architecture, suggesting that with increased model complexity or rigorous hyperparameter tuning, it has the capacity to outperform the ResNet model.

Additionally, our venture into exploring Vision Transformers represented a significant expansion of our research horizon. The outcomes revealed promising correlations among dimensions and a notable Average Correlation Coefficient. These findings underscore the potency of Vision Transformers, hinting at their potential for substantial advancements. The encouraging correlations across dimensions and the substantial Average Correlation Coefficient serve as catalysts, urging us to delve deeper into refining these models further. There's a clear indication that dedicating more effort to these models could significantly enhance their performance in deriving Multidimensional Scaling (MDS) Values.

Table 5. Correlation Coefficient Across Dimensions

Dimension	Characteristics	360 Image set	120 Image Set
1	Lightness	0.6351	0.6596
2	Avg. Grain Size	0.4966	0.4261
3	Roughness	0.2181	0.4027
4	Shininess	0.3384	0.3439
5	Organization	0.1954	0.3637
6	Chromaticity	0.5005	0.6282
7	Shape	0.3534	0.2231
8	Hue	0.4216	0.3141

Despite the time constraints, our exploration into these architectures—DenseNet and Vision Transformers—has provided valuable insights into their capabilities and potential. The observed results not only validate the effectiveness of DenseNet but also highlight the promising avenues for improvement and refinement in Vision Transformers, motivating us to invest more resources and effort into these transformative models for enhanced performance in deriving MDS values.

4. Discussion

In the course of this project, it is crucial to address the following inquiries to comprehend the project's purpose, its essentiality, the rationale behind method selection, and the choice of specific datasets.

1. What is the reason behind utilizing a limited dataset for our neural network, specifically comprising only 360 rock images?

To address the question regarding the modest size of our dataset (comprising only 360 rock samples), it is crucial to revisit the primary goal of this project. We have adopted the Generalized Context Model (GCM)(*Nosofsky et al. 1986* [9], *2011* [1]0), a widely recognized psychological framework employed by cognitive psychologists to elucidate the processes of human object categorization and judgment formation based on similarity and feature overlap.

The GCM operates by utilizing a higher-dimensional feature space as its input, encapsulating object characteristics through the quantification of diverse attributes. In the context of rock classification, dimensions within this feature space may encompass properties like porosity and smoothness. The creation of this feature space is facilitated by Multidimensional Scaling (MDS), a technique that transforms a matrix of similarity or dissimilarity values into a higher-dimensional representation.

Generating MDS values for a set of n objects necessitates knowledge of n(n-1)/2 similarity/dissimilarity values. Given our specific focus on understanding how humans classify objects, collecting such data from human subjects for the 64620 data cells corresponding to our dataset of 360 rock images is both time- and resource-intensive. Consequently, the MDS space for this dataset is derived from a similarity matrix, where a considerable number of cells are based on limited observations, and several cells remain entirely empty.

This constraint leads to the utilization of a smaller dataset of 360 rock images. However, our strategic approach involves training a neural network to accurately generate MDS coordinates for given objects. Once this task is accomplished, we can seamlessly incorporate more objects into the dataset. The advantage of employing neural networks lies in their capacity for automatic embedding in the psychological space, enabling the inclusion of an unlimited number of additional objects from the relevant category domain.

It is important to note that the overarching aim of integrating neural networks into our methodology is to produce MDS coordinates for real-world objects, thereby facilitating the expansion of our dataset with diverse items from natural categories. These enriched datasets can subsequently serve as inputs to psychological models like the GCM, allowing us to gain deeper insights into the intricacies of human categorization.

2. Why are we training the neural network to generate embeddings for rocks(what is the need to do so)?

It is crucial to emphasize that the project's primary goal is to gain insights into how humans categorize objects, presenting a challenge beyond straightforward classification. The psychological models employed by cognitive psychologists in this context have predominantly been tested on artificial object categories, such as geometric forms, distinguished by simple features like shape, size, and color. However, a more robust evaluation of these models involves using real-world objects, where the complexity of features necessitates a more intricate approach to object categorization.

We opted to focus on rocks as the subject of their experiment for several compelling reasons. Rocks, being natural stimuli, offer a complexity of psychological dimensions that is often challenging to articulate or quantify through traditional methods. The graded structures within rock categories, with prototypical and less typical members, as well as the notable within-category variability, make rock classification a representative example of natural category learning.

The challenges posed by the overlapping category distributions, fuzzy boundaries, and the need to integrate information across complex, high-dimensional feature spaces aligns with our objective of developing a method capable of handling real-world categorization intricacies. Furthermore, the experiment benefits from the relative lack of detailed prior knowledge among participants regarding the structure of rock categories in the geologic sciences, allowing for precise experimental control in the laboratory setting.

Importantly, it is believed that the proposed method, integrating traditional psychological scaling techniques and deep-learning networks, has the potential to be applied across a broad spectrum of naturalistic domains, contributing to the advancement of computational models in cognition and behavior.

3.Why is it important for psychologists to comprehend the process of human object categorization?

Understanding how humans classify objects holds practical significance for cognitive psychologists across various domains. It plays a pivotal role in refining artificial intelligence and machine learning algorithms, improving image recognition, and supporting autonomous systems. This knowledge contributes to the design of more intuitive interfaces, enhancing user experience. In marketing, it informs strategies for product design and advertising, while in education, it guides curriculum development and teaching methodologies. Overall, insights into human object classification have far-reaching applications, influencing technology, design, marketing, and education.

5. Future Work

1. Investigate advanced architectures and optimization strategies to enhance neural network performance in generating Multidimensional Scaling (MDS) solutions.

2. Explore techniques, beyond traditional similarity judgments, for data augmentation and noise reduction in the MDS space to improve the quality of training data for CNN-derived MDS coordinates.

3. Investigate methods to improve the interpretability of deep learning representations, especially hidden-layer activations, for more meaningful insights into learned features.

4. Explore the feasibility and effectiveness of simultaneous training on both similarity-judgment and classification data to discover a more comprehensive set of psychologically relevant dimensions.

5. In terms of data, the MDS coordinates associated with the 360-rock dataset have been obtained from a similarity matrix characterized by restricted observations and numerous incomplete cells. This incompleteness introduces noise into the MDS feature space. To address this, enhancing the dataset by collecting additional similarity judgments and filling the missing entries in the 360×360 similarity matrix, used as input for Multidimensional Scaling (MDS), can effectively mitigate this noise.

6. Related work

1. *Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017)*: Explores natural-science categories deviating from the family-resemblance principle, offering insights into category learning. https://doi.org/10.1177/0956797616675636

2. *Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018a)*: Introduces a psychological model of classification applied to natural-science category learning. https://doi.org/10.1177/0963721417740954

3. *Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018b)*: Examines an exemplar-memory model of classification learning in a high-dimensional natural-science category domain.
https://doi.org/10.1037/xge0000369

4. *Austerweil, J. L., & Griffiths, T. L. (2011)*: Investigates the impact of distributional information on feature learning, contributing to theoretical advancements in human feature learning.

5. *Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2017)*: Explores human categorization of natural images using deep feature representations, bridging cognitive science and deep learning.

6. *Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2019)*: Extends prior work by combining deep networks and cognitive models to capture human categorization behavior at a larger scale.

7. *Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017)*: Investigates deep neural networks' performance compared to humans in object recognition tasks under degraded signal conditions, uncovering insights into model robustness and limitations.

8. *Guest, O., & Love, B. C. (2017)*: Explores brain imaging's implications on understanding the neural code, linking neural representations to observed behavior.

9. *Jacobs, R. A. & Bates, C. J. (2019)*: Compares visual representations and performance between human observers and deep neural networks, highlighting similarities and differences in visual information processing.

10. *Nasr, K., Viswanathan, P., & Nieder, A. (2019)*: Investigates the emergence of number detectors in deep neural networks designed for visual object recognition, unveiling aspects related to numerical cognition.

11. *Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018)*: Conducts a large-scale comparison of human, primate, and deep neural network object recognition behaviors, elucidating their similarities and differences.

All these related works contribute significantly to understanding object recognition, human categorization, and the interplay between deep neural networks and human visual processing. They provide crucial insights in understanding the complex interplay between computational models and human perceptual processes that inform our project's exploration into constructing psychological feature spaces for object categorization using alternative neural network architectures.

References

[1] Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018c): Toward the development of a feature-space representation for a complex natural category domain. Behavior Research Methods, 50(2), 530–556. https://doi.org/10.3758/s13428-017-0884-8.

[2] Nosofsky, R. M., Sanders, C. A., Meagher, B. J., Douglas, B. J. (2019a): Search for the missing dimensions: building a feature-space representation for a natural-science category domain. Computational Brain & Behavior, 1–21

[3] Sanders, C.A., Nosofsky, R.M. (2020): Training Deep Networks to Construct a Psychological Feature Space for a Natural-Object Category Domain. Comput Brain Behav 3, 229–251 (2020).
https://doi.org/10.1007/s42113-020-00073-z

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2015): ”Deep residual learning for image recognition”. arXiv preprint arXiv:1512.03385.

[5] Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (2019b): ”Model-guided search for optimal natural-science-category training exemplars: a work in progress”. Psychonomic Bulletin & Review, 26(1), 48–76.

[6] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017): ”Densely connected convolutional networks”. arXiv preprint arXiv:1608.06993.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020): ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. arXiv preprint arXiv:2010.11929 [cs.CV].

[8] Wu, B., Xu, C., Dai, X., et al. (2020): ”Visual Transformers: Token-based Image Representation and Processing for Computer Vision”. arXiv preprint arXiv:2006.03677 [cs.CV]

[9] Nosofsky, R. M. (1986): ”Attention, similarity, and the identification-categorization relationship”. Journal of Experimental Psychology. General, 115(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39.

[10] Nosofsky, R. M. (2011): ”The generalized context model: an exemplar model of classification”. In Pothos, E. M. and Wills, A. J. (Eds.), Formal approaches in categorization, 18–39. Cambridge University Press.