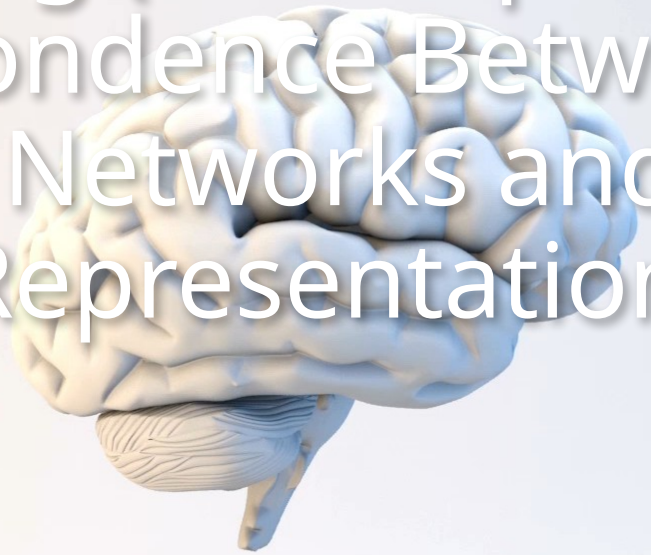
Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations



Joshua C. Peterson, Joshua T. Abbott, Thovmas L. Griffiths (P.A.G)

(w/ graphics from Ayşe Aybüke Durmaz, Homa Priya Tarigopula)

P.A.G Project Overview

Introduction

Aims

Evaluating the correspondence between representations

Stimuli,
Methods,
Results,
Analysis

Reweighting starts like RSA

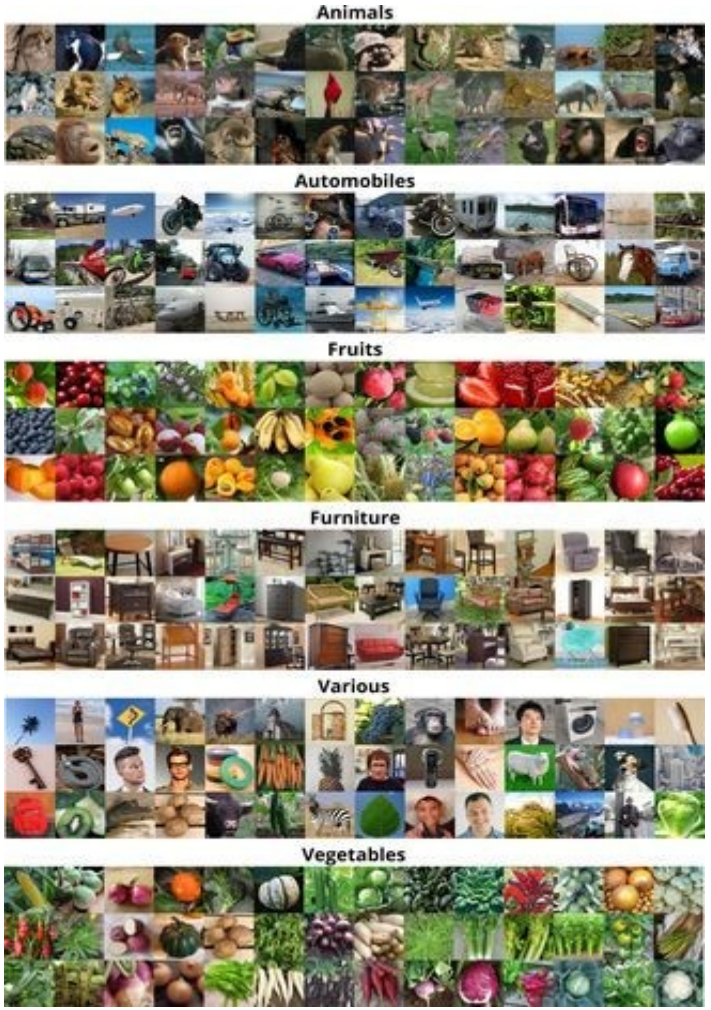
- Comparing the representations formed by deep neural networks to those of humans.
- Human representation:
  - Similarity judgments.
  - A similarity function over a set of pairs of data points corresponds to an implicit representation of those points

Evaluating the correspondence between representations

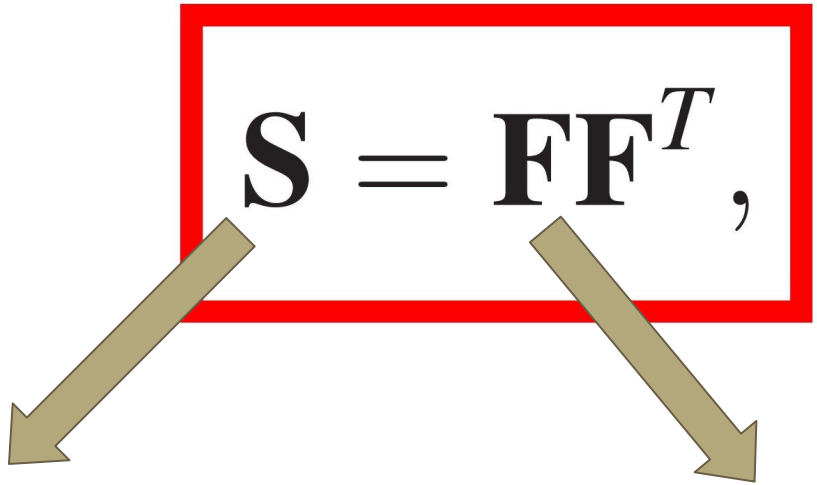
- Obtain the similarity within images using human ratings
- Find the similarity within images of each category by
  - Studying Raw representations of DNNs
  - Studying Transformed Representations of DNNs
- Probe both DNN and human similarities to identify the spatial and taxonomic information encoded
  - MDS, HCA

Evaluating the correspondence between representations

- Stimuli
  - 6 categories - 120 images each
- Procedures:
  - pairwise image similarity ratings
  - Amazon Mechanical Turk
    - 0 (“not similar at all”)
    - 10 (“very similar”)
  - six 120x120 similarity matrices



Evaluating the correspondence between representations



$$A = [x0, x1, x2, \ldots, xn]$$

$$B = [y0, y1, y2, \ldots, yn]$$

$$A \odot B = \sum_{i=0}^{n} xi \cdot yi$$

SIMILARITY MATRIX from humans

FEATURE MATRIX (dot product/ R/ cosine)

Evaluating the correspondence between representations

$$\mathbf{S} = \mathbf{F}\mathbf{F}^T,$$

FEATURE MATRIX

Deep neural network representations :

- When deep neural networks are presented with an image, the nodes that comprise the network obtain different activation values.

Multidimensional feature representation

120 images

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Evaluating the correspondence between representations

$$\mathbf{S} = \mathbf{F}\mathbf{F}^T,$$

FEATURE MATRIX

120 images

Multidimensional feature representation [ncol :4,096]

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Evaluating the correspondence between representations

$$\mathbf{S} = \mathbf{F}\mathbf{F}^T,$$

FEATURE MATRIX

Inner Product of Feature Matrix produces a 120x120 similarity matrix:

F = [120 x 4096]

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

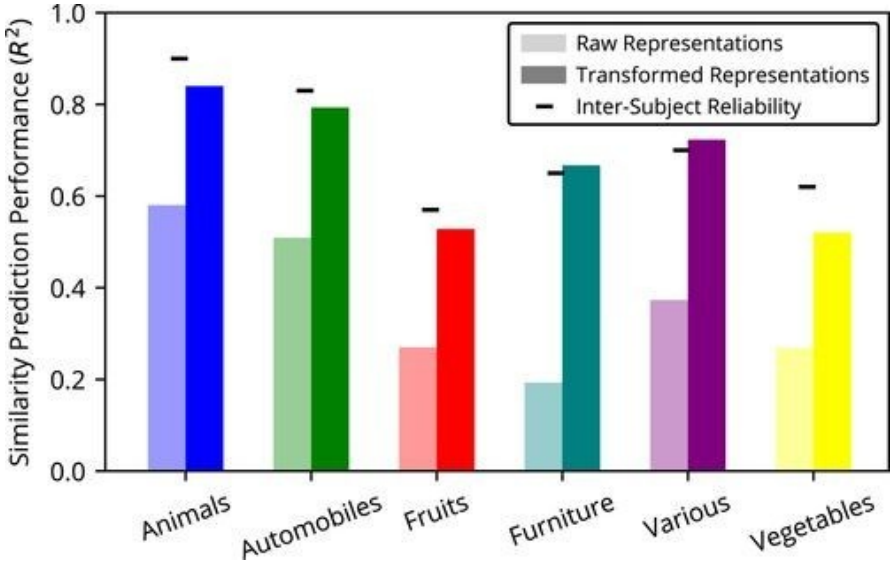
F^T = [4096 x 120]

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Evaluating the correspondence between representations

Results :

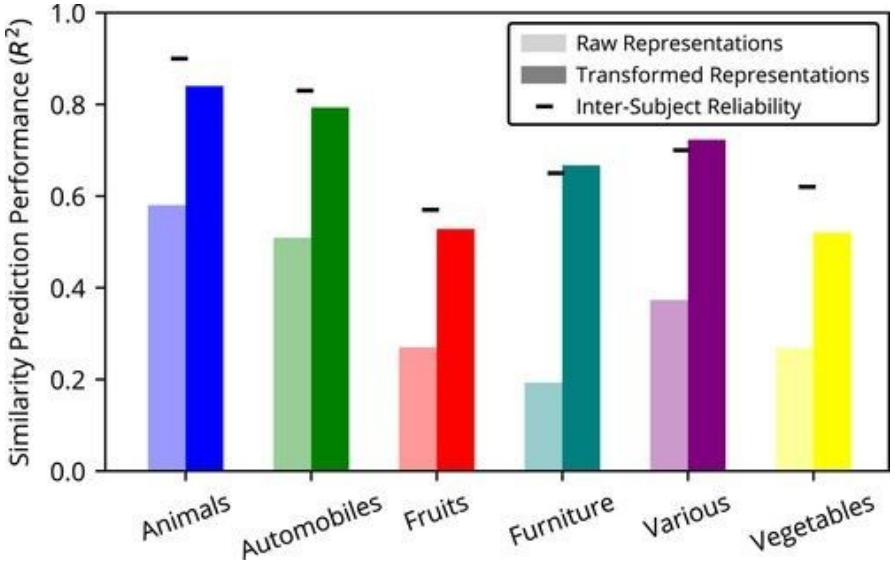
- computing the correlation between the human similarity judgments and the inner products computed in the deep feature spaces. (both 120x120)



Evaluating the correspondence between representations

Results :

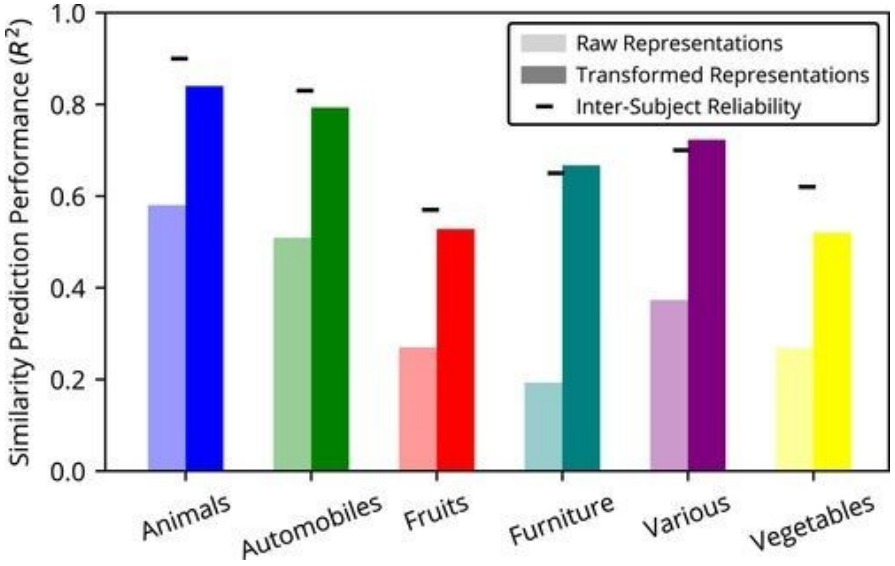
- computing the correlation between the human similarity judgments and the inner products computed in the deep feature spaces. (both 120x120)
- VGG (best performing DNN architecture)



Evaluating the correspondence between representations

Results :

- computing the correlation between the human similarity judgments and the inner products computed in the deep feature spaces. (both 120x120)
- VGG (best performing DNN architecture)
- The raw deep representations provide a reasonable first approximation to human similarity judgments

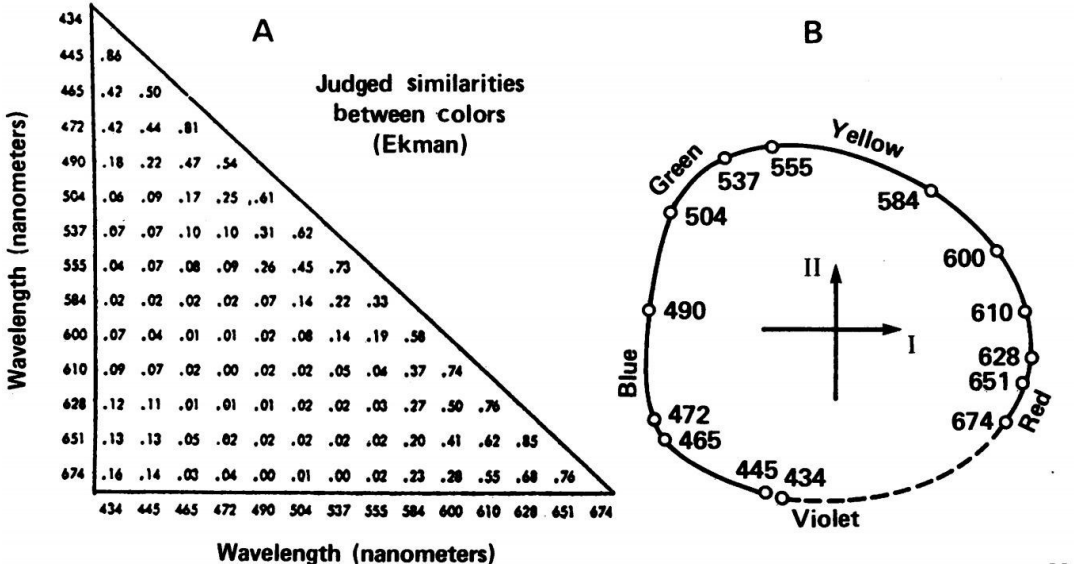


Evaluating the correspondence between representations

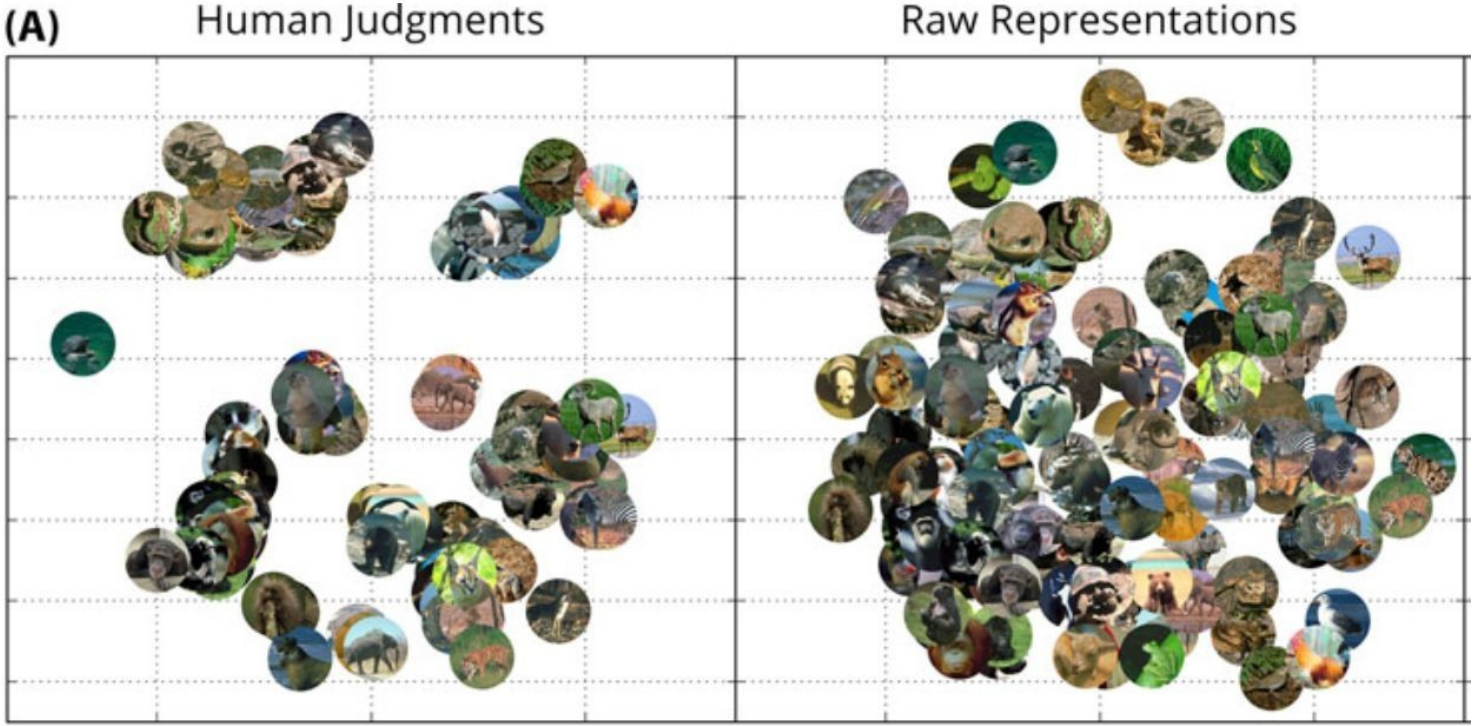
To better understand how DNNs succeed and fail to reproduce the structure of psychological representations, they applied two classic psychological tools:

- non-metric multidimensional scaling
  - converts similarities into a spatial representation
- hierarchical clustering
  - produces a tree structure (dendrogram)

Multidimensional Scaling: from relative distances of n elements to a map of relative distances in 2* dimensions.

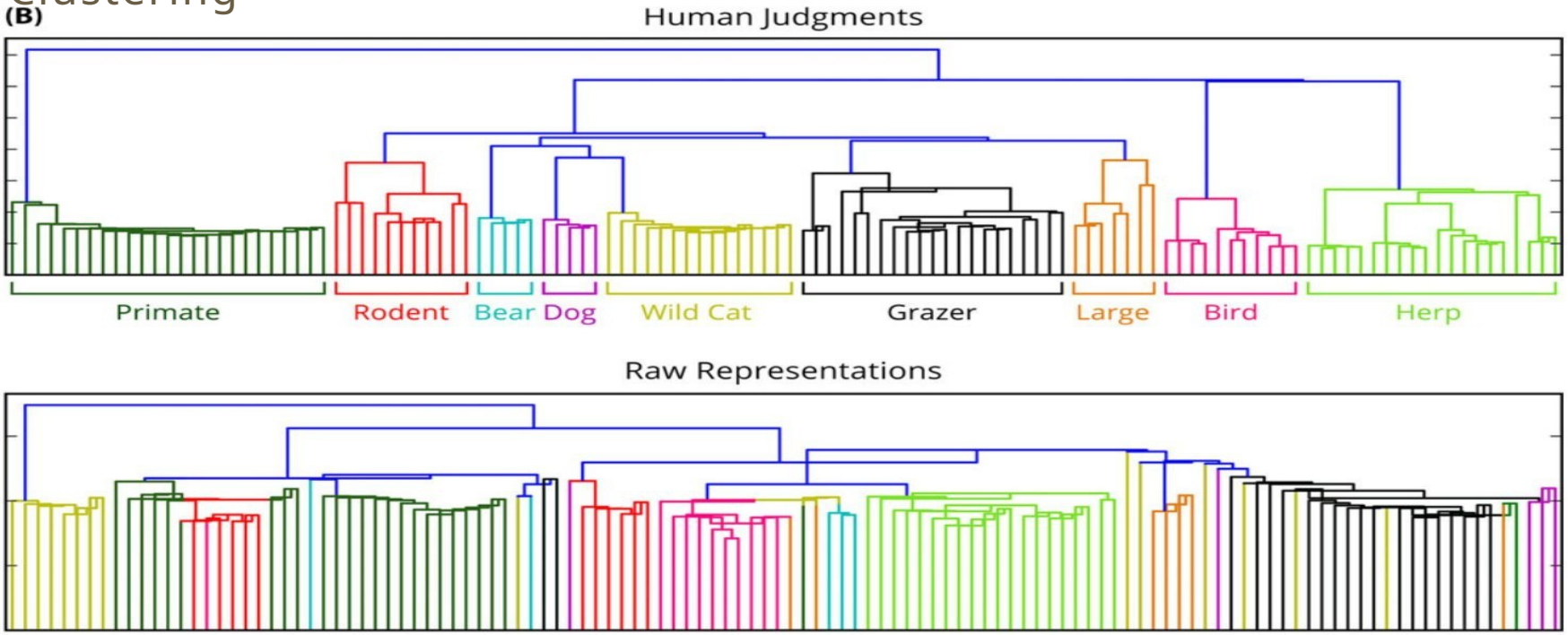


Evaluating the correspondence between representations

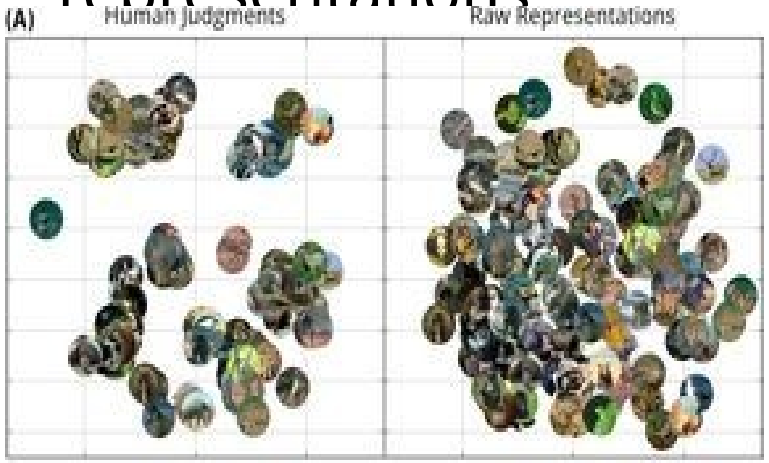


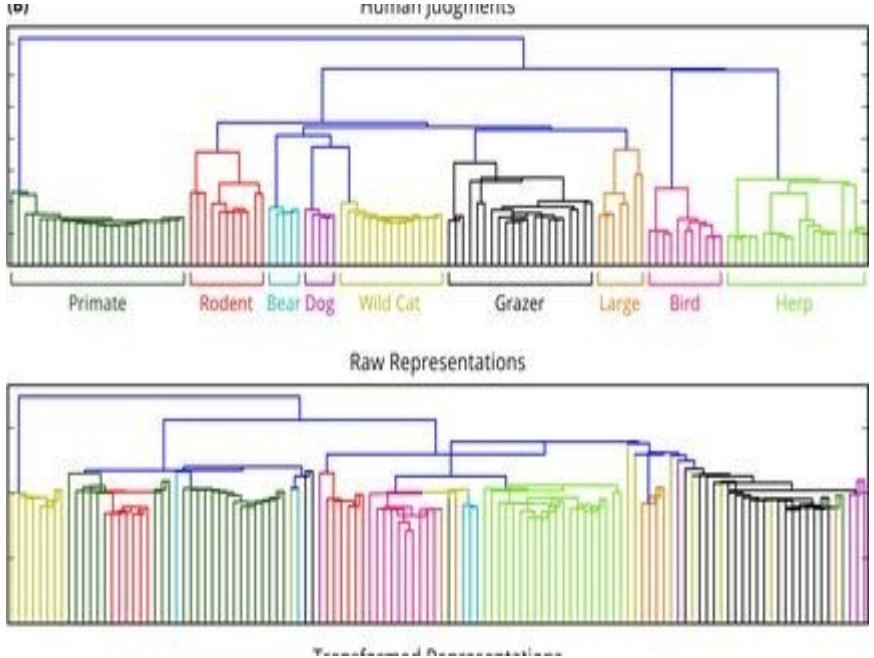
Evaluating the correspondence between representations

2. Hierarchical Clustering



Evaluating the correspondence between representations





Human representations exhibit highly distinguished clusters in the spatial projections and intuitive taxonomic structure in the dendrograms, neither of which is present in the DNN representations.

Transforming Representations

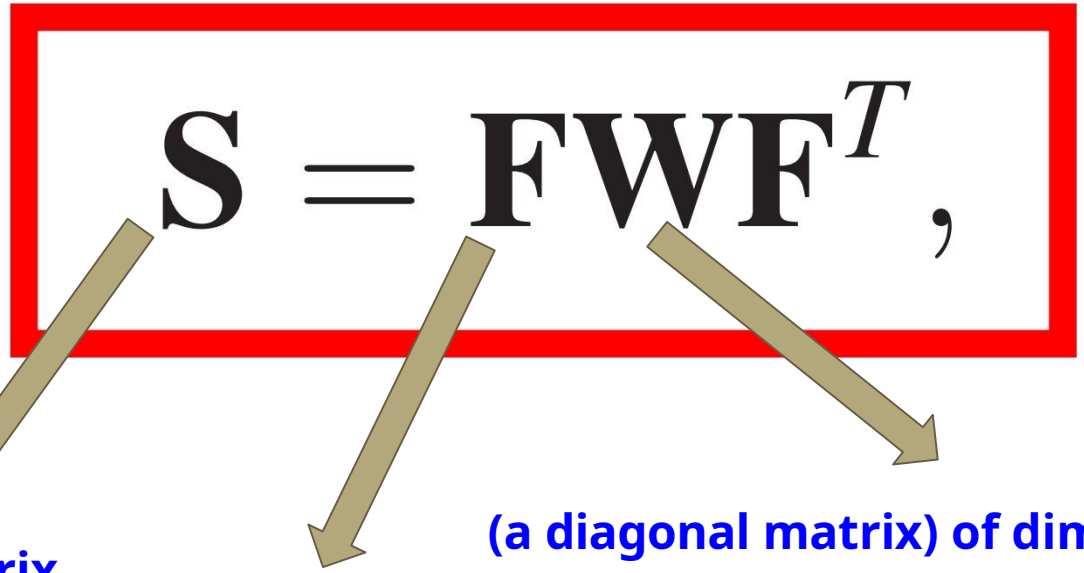
- how can DNN representations be transformed to increase their alignment with psychological representations?
  - with a set of weights on the features used to compute similarity

Transforming Representations

- how can DNN representations be transformed to increase their alignment with psychological representations?
  - with a set of weights on the features used to compute similarity

$$\mathbf{S} = \mathbf{F}\mathbf{W}\mathbf{F}^{T},$$

Transforming Representations



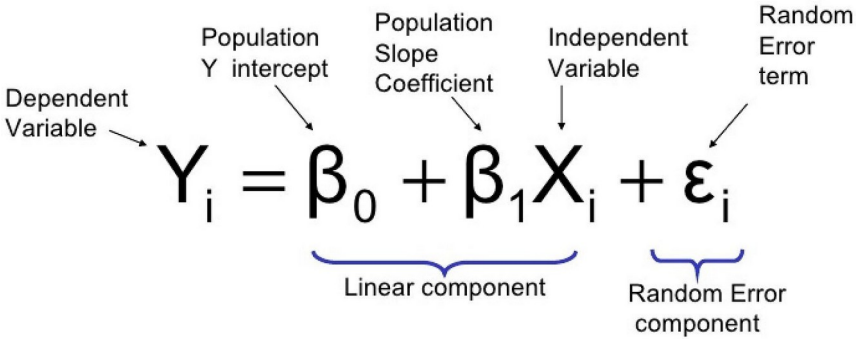
Transforming Representations

- It provides a way to specify the relationship between a feature representation and stimulus similarities.

$$\mathbf{S} = \mathbf{F}\mathbf{W}\mathbf{F}^{T},$$

Transforming Representations

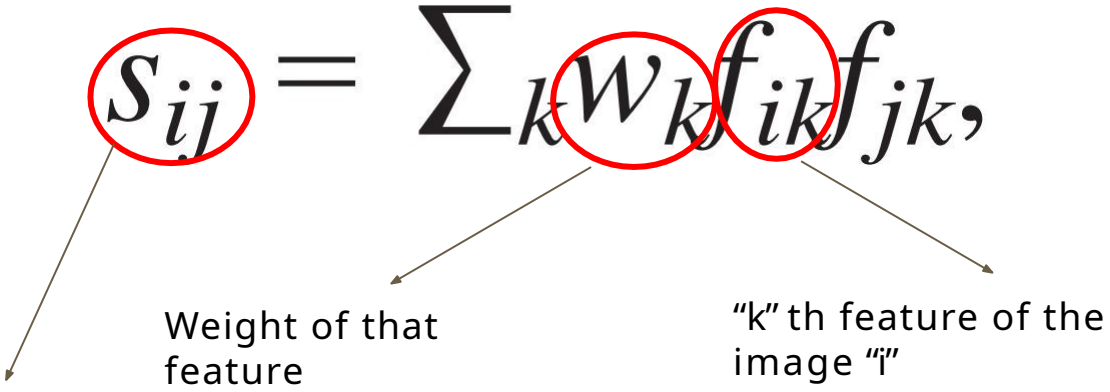
- can be expressed as the **solution to a linear regression problem**
- the predictors for each similarity **s***ij* are the (elementwise) product of the values of each feature for objects *i* and *j*
- (i.e., each row of the regression design matrix **X** can be written as **Fi∘Fj**)



$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \dots x_{1,\,p-1} \\ x_{20} & x_{21} & \dots x_{2,\,p-1} \\ \dots & \dots & \dots \quad \dots \\ x_n & x_{n1} & \dots x_{n,\,p-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Transforming Representations

- The similarity **s***ij* between objects *i* and *j* is therefore modeled as :



Learning transformations

- Freely identifying the **w** that best predicts human similarity judgments runs the risk of overfitting, since our DNNs generate thousands of features.

Learning transformations

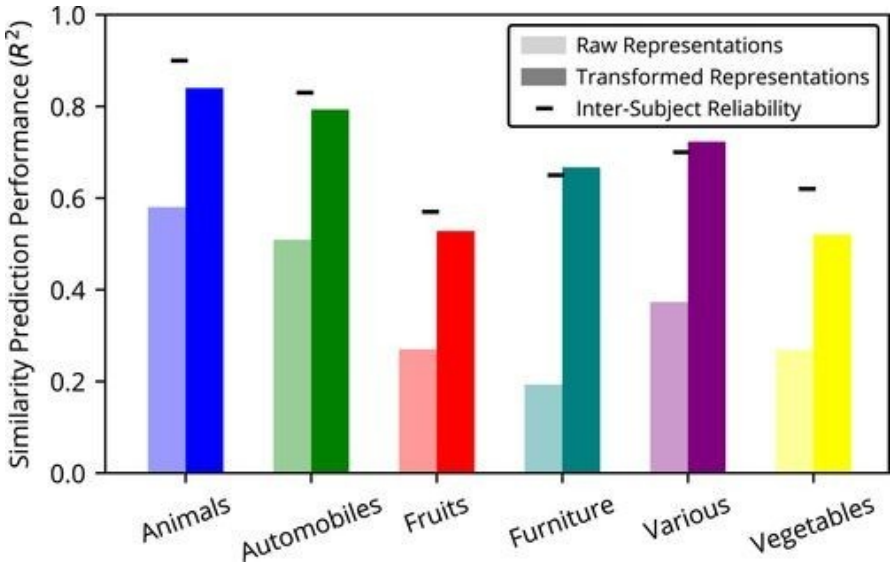
- Freely identifying the **w** that best predicts human similarity judgments runs the risk of overfitting, since our DNNs generate thousands of features.
- L2 regularization on **w**
  - penalizing models for which the inner product $w^T w$ is large

Learning transformations

- Freely identifying the **w** that best predicts human similarity judgments runs the risk of overfitting, since our DNNs generate thousands of features.
- L2 regularization on **w**
  - penalizing models for which the inner product $w^T w$ is large (we want to minimize the square magnitude of weights; produces models with may small weights rather than few large ones.)

Improvement through feature adaptation

- Trained only for the best performing Deep Neural Network, **VGG**
- Variance Explained **doubled** compared to raw representations on all domains



Improvement through feature adaptation

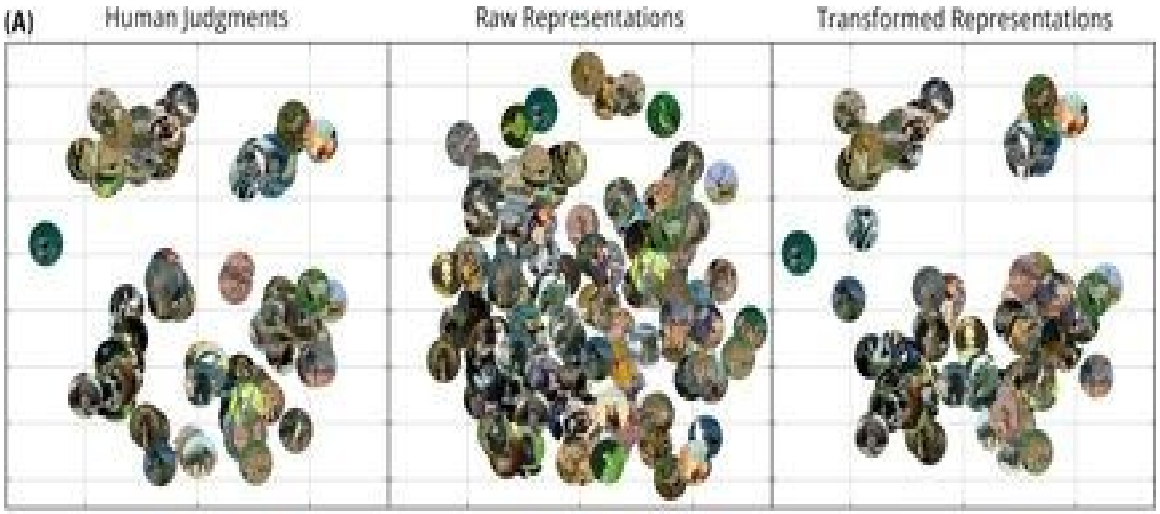
- The stricter case of cross-validation titled ‘CV control’ explained the similarity judgements better than raw representations. In this case no single images occured in both training folds and test folds of cross-validation
- However, for ‘Transformed model’, the exclusivity was wrt to pair of images.

Table 1. Variance explained in human similarity judgments for raw and transformed representations for the best performing network (VGG)

Dataset	Raw R^2	Transformed R^2	CV Control R^2	Human Inter-reliability
Animals	0.58	0.84	0.74	0.90
Automobiles	0.51	0.79	0.58	0.83
Fruits	0.27	0.53	0.36	0.57
Furniture	0.19	0.67	0.35	0.65
Various	0.37	0.72	0.54	0.70
Vegetables	0.27	0.52	0.35	0.62

Improvement through feature adaptation

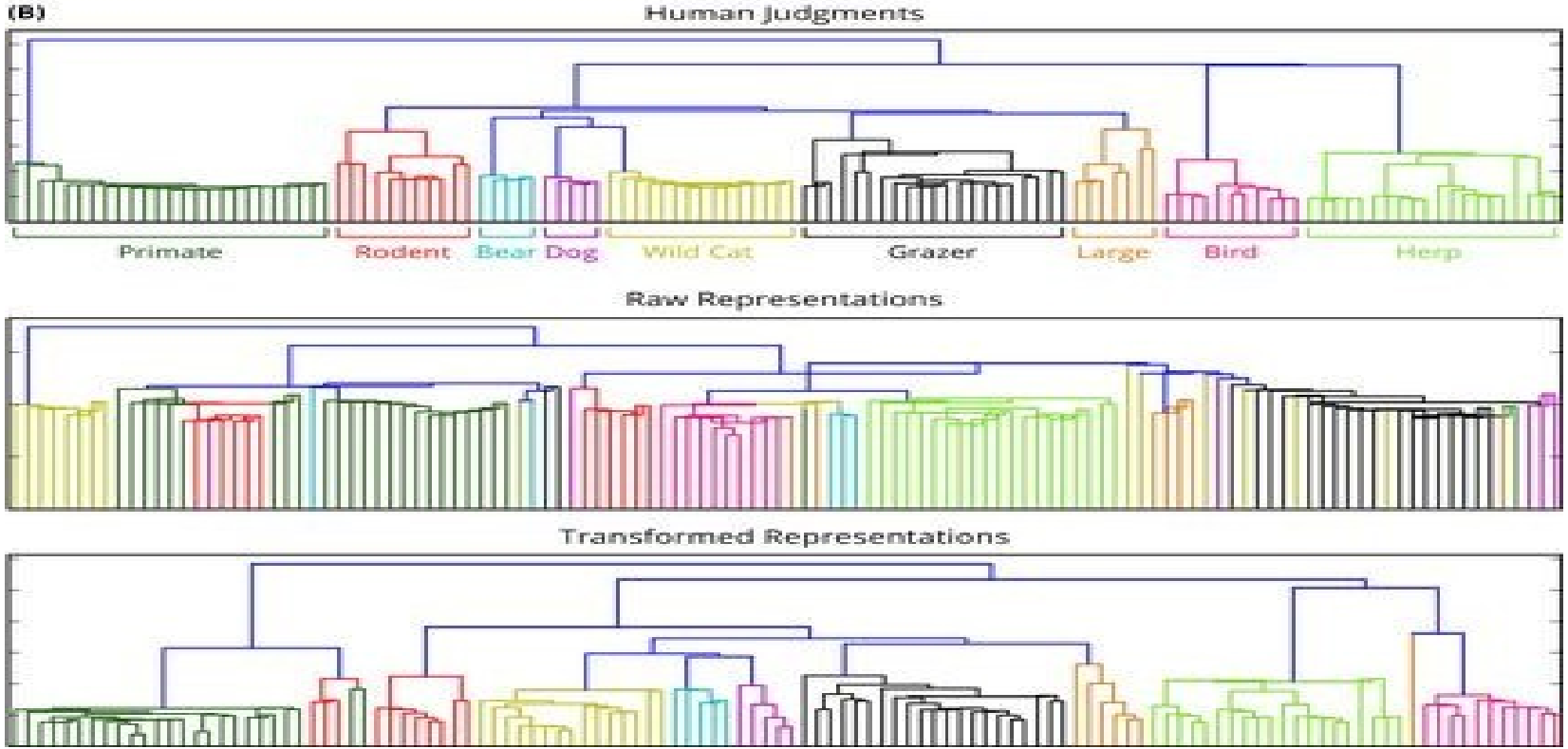
Comparison of the results from MDS :



Stronger resemblance to

human judgements

Improvement through feature adaptation

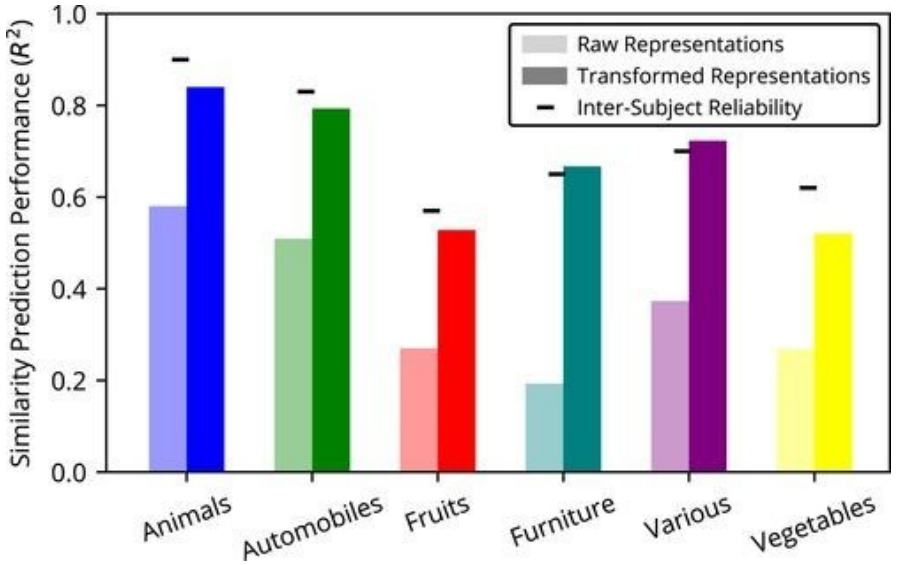


Inter-domain Transfer

- Transformations learnt on one domain do not generalize to others
- Correlations are poor & worse compared to those of

Inter-domain generalization of best performing DNN transformations

Training Set	Test Set	R^2
Animals	Fruits	0.11
Animals	Furniture	0.02
Animals	Vegetables	0.11
Animals	Automobiles	0.17
Animals	Various	0.12



Discussion

- Raw representations from an AI system may approximate well representational space of humans, but MDS and Hierarchy show weakness.
- These representations can be easily transformed to produce better human similarity predictions
- These transformed models can be used to predict similarity judgements for new stimuli

Discussion

- The simple re-weighting of features that the linear transformation performs, can be viewed as an analogue to dimensional attention.
- The ability of transformed representations to generalise for new stimuli empowers studies on cognitive processes relying on such representations in the brain.