Understanding Human and AI Representations Through Mutual Constraints

ABNS 2023/2024.

Organization of semantic domains

Psychology: Semantic domains or categories (e.g., mammals, animals, dogs) organized via features or dimensions that carry the relevant variance for the category. Refresh: Rosch et al. Lake et al. Classical view.

AI: entities (e.g., images, words) described by feature-values from which representational category effects emerge. Refresh: Lake et al.

AI systems as models of semantics

AI systems trained for image categorization or word embedding produce representations that reasonably approximate those of humans:

- Similarity between categories, as operationalized from human data, is well predicted by distances between objects in the AI model.
  - Human similarity is quantified using brain/behavior.
  - Model similarity is quantified via Euclidean distances, inverse Cosine, etc'.
  - Refresh: Kriegeskorte et al. 2008.

Representational Similarity Analysis

Vs. human Similarity judgments:

R^2 ~ 0.2 – 0.6; Peterson J.C., Abbott J.T., Griffiths T.L. 2018

R = 0.56; King M.L., Groen I.I., Steel A., Kravitz D.J., Baker C.I. 2019

R = 0.26; Groen I.I., Greene M.R., Baldassano C., Fei-Fei L., Beck D.M., Baker C.I. 2017

Vs MVPA brain activity (**see last lecture**)

R ~ 0.04 – 0.07 for various brain areas; Liuzzi A.g., Aglinskas, A., Fairhal S.L. 2020

R < 0.2; Groen et al.

AI modeling of human representations

Approach 1 (Default): Use of *all* DNN features as object-representation for modeling human data.
Implicitly assumes : All features are in relevant and equally important for all concepts.

Approach 2 (Reweighting). Addresses mis-calibrated, human-relevant features.
Assumes: AI learns human-relevant features, but these are mis-calibrated. Applies concept-specific adjustment of feature saliency for modeling human representations.

Approach 3 (Pruning). Investigates modular structures in AI models.
Assumes: The network develops a modular structure where information about different categories is represented in different subspaces (subsets of latent dimensions) within the model.

Why important?

Psychology: AI models achieve human-like competence on different tasks; a model of potential knowledge organization. And interpretability.

Engineering: Better prediction of human behavior; improved AI-human alignment.

Computer Science: Understanding representations in neural networks.

Status as models: Approach 1; use-as-is

Approach 1 (Default): Use *all* DNN features as object-representation for modeling human data.

Assumes: All features are relevant and equally so for all concepts.

Assuming two objects, *U*, *V*, each with 3 features, human similarity can be defined as inner product (or related quantity).

$$Similarity = V \cdot U = V_1 \cdot U_1 + V_2 \cdot U_2 + V_3 \cdot U_3$$

Status as models: Approach 2; use all but reweight

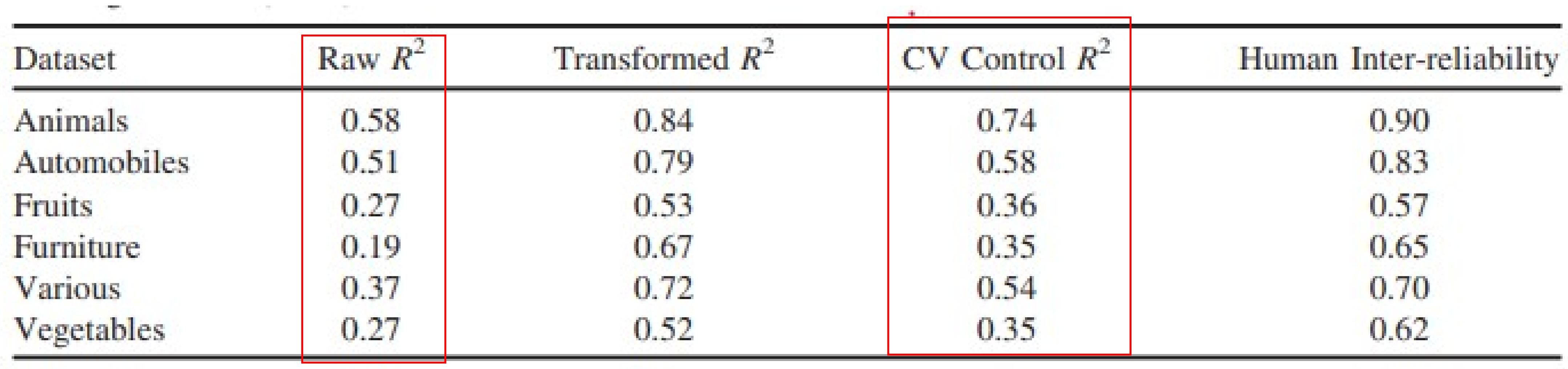
Approach 2 (reweighting): learn a set of weights that modifies the saliency of each feature so that the inner dot product better predicts human similarity.

Implemented post-training.

Assuming two objects, *U*, *V*, each with 3 features, similarity is defined as weighted inner product or related measure. Weights (W_{123}) learned via regression; evaluated on out-of-sample data.

$$Similarity(V,U) = (V_1 \cdot W_1 \cdot U_1) + (V_2 \cdot W_2 \cdot U_2) + (V_3 \cdot W_3 \cdot U_3)$$

Reweighting is effective



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

Peterson et al. 2018

Status as models: Approach 3; select without reweighting

Approach 3 (Pruning) : Assume that diverse human categories are captured by different latent dimensions in the DNN’s features space.

Aim: identify a subset of features, per category, that capture these dimensions, such that **using this subset alone,** improves prediction of human representations (vs. full feature-set).

Reweighting vs. pruning.

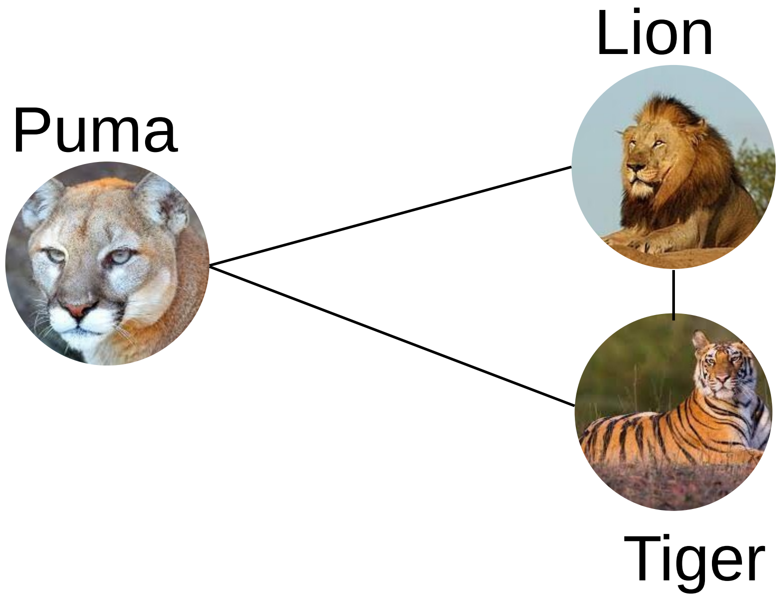
$$Similarity_{full}\left(V,U\right) = \left(V_1 \cdot W_1 \cdot U_1\right) + \left(V_2 \cdot W_2 \cdot U_2\right) + \left(V_3 \cdot W_3 \cdot U_3\right)$$

$$Similarity_{pruned}\left(V,U\right) = \left(V_2 \cdot U_2\right) + \left(V_3 \cdot U_3\right)$$

How pruning works

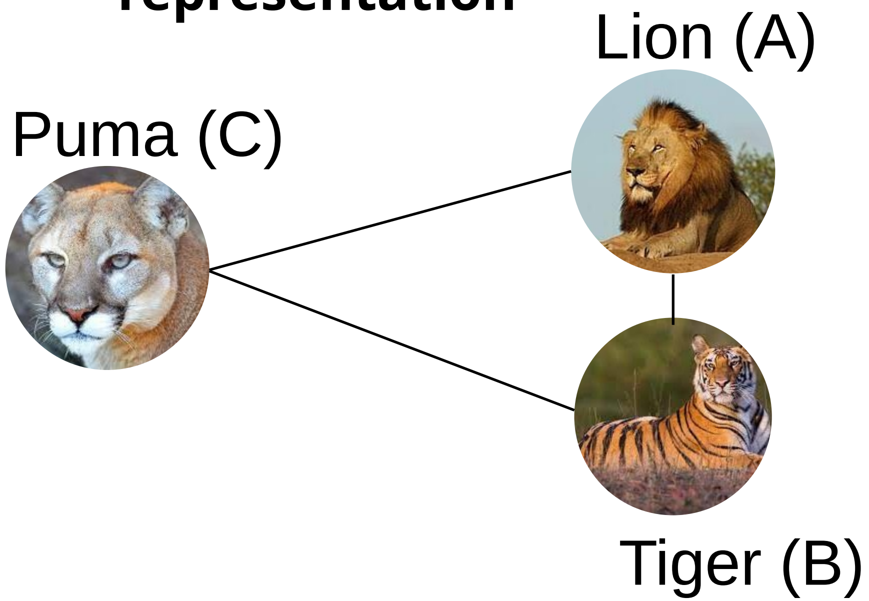
Pruning aims to improve on initial match between human and model representations.

Toy example: Humans find Tigers more similar to Lions than to Pumas



Pruning works by improving on initial match

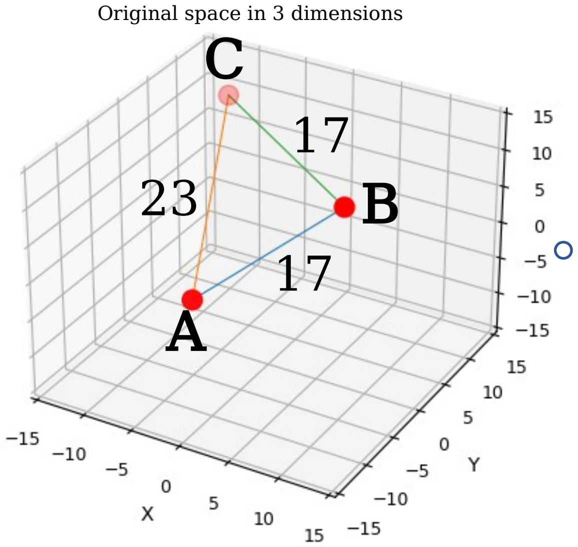
Human representation



Model representation

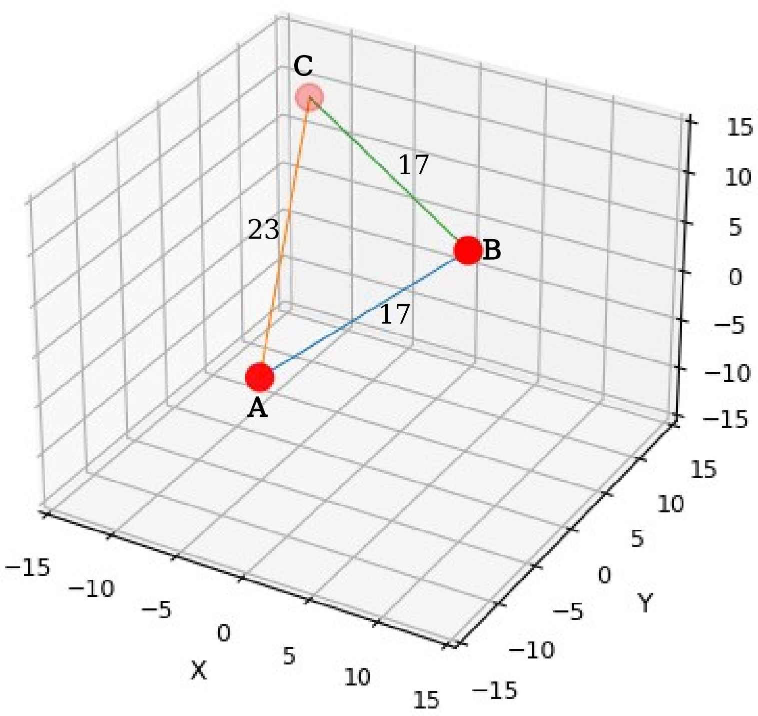
Feature values (image embeddings)

	Feature1	Feature2	Feature3
Lion (A)	-5	-5	-5
Tiger (B)	5	5	5
Puma (C)	-10	10	12

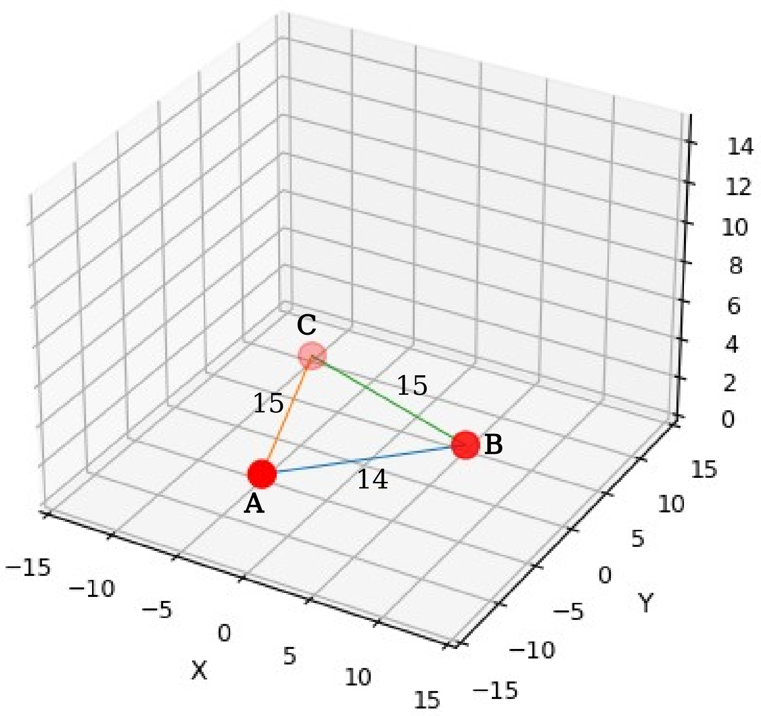


On full model,
Sim(BA) =
Sim(BC)

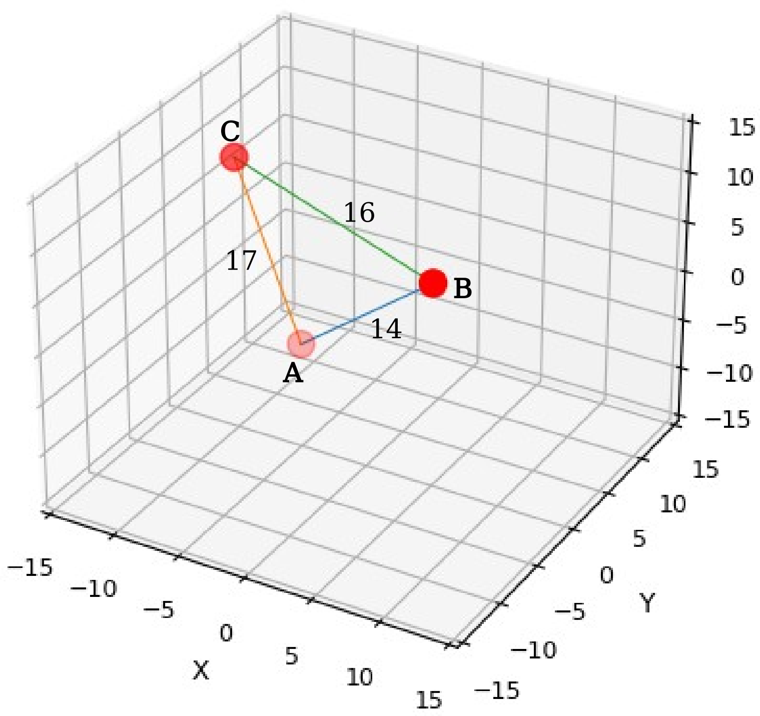
Original Data



Z-Dimension (f3) Removed



Y-Dimension (f2) Removed



	f1	f2	f3
Lion (A)	-5	-5	-5
Tiger (B)	5	5	5
Puma (C)	-10	10	12

	f1	f2	f3
Lion (A)	-5	-5	~~-5~~
Tiger (B)	5	5	~~5~~
Puma (C)	-10	10	12

	f1	f2	f3
Lion (A)	-5	~~-5~~	-5
Tiger (B)	5	~~5~~	5
Puma (C)	-10	10	12

By iterating over features and using *Sequential Feature Selection* algorithms, supervised pruning learns a subset of features that better predicts human judgments and generalizes to out of sample data.

In what follows; summary of pruning to improve representational similarity between AI and Humans

1. Pruning
   1. Improves out-of-sample prediction accuracy for human similarity judgments of images, (higher RSA isomorphism).
   2. Produces a more psychologically valid representational space
   3. Improves prediction of out-of-sample MVPA data (Brain RDMs).
2. The feature-sets retained by pruning vary depending on the category guiding the pruning process, and these sets identify different subspaces (latent factors) in the feature space.
3. Pruning improves out of sample prediction of human similarity judgments for words and allows an interpretation of latent dimensions underlying word similarity judgments.

Pruning improves prediction of human similarity judgments

Study: learning to predict human similarity judgments within 6 categories, each consisting of 120 images.

Model pruned: penultimate layer of VGG19, with 4096 nodes (features).

```
Improved prediction of
behavioral and neural
similarity spaces using pruned
DNNs.
Tarigopula P, Fairhall SL,
Bavaresco A, Truong N, Hasson
U.
```

Neural Networks, 168:89-104,

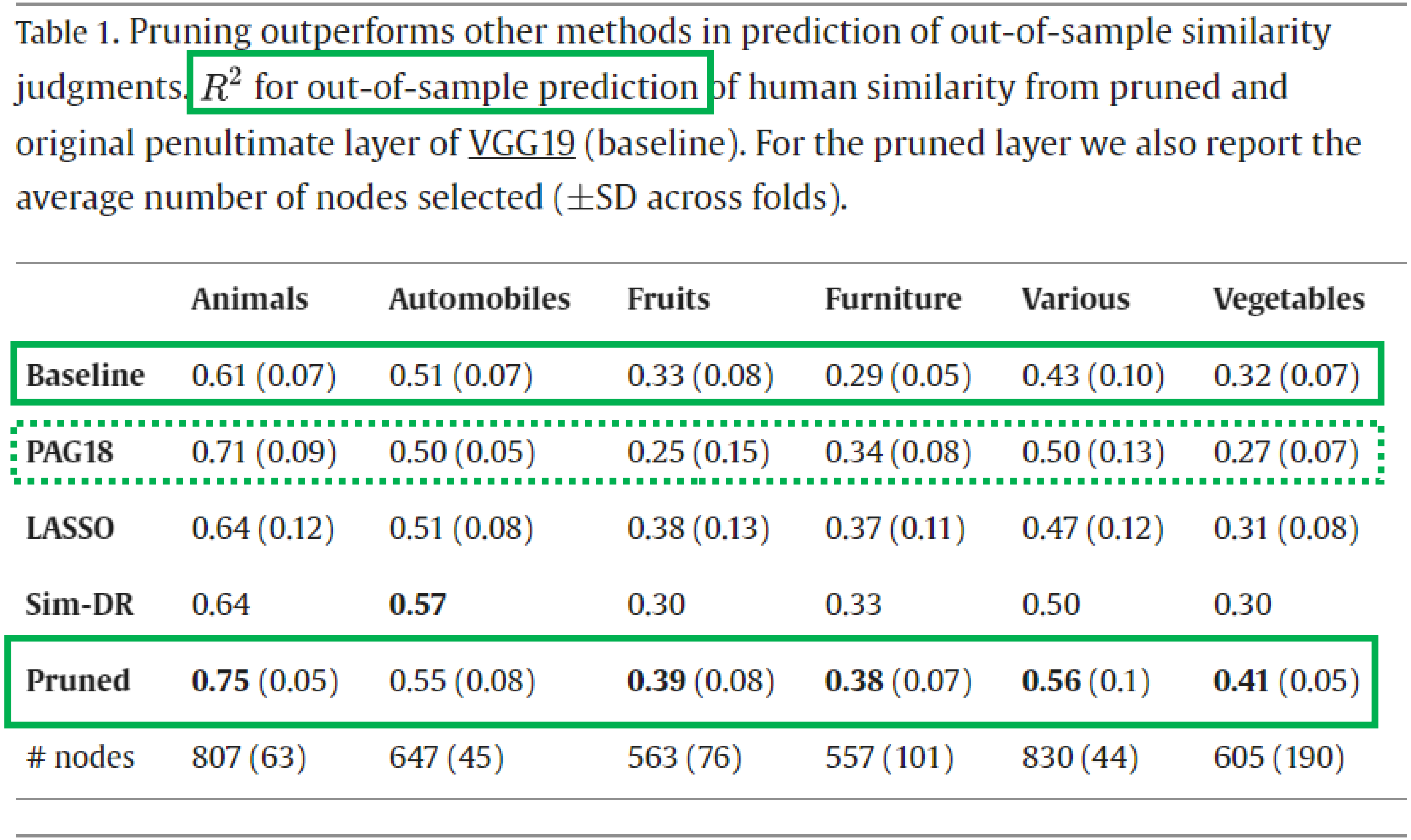


Table 1. Pruning outperforms other methods in prediction of out-of-sample similarity judgments. R^2 for out-of-sample prediction of human similarity from pruned and original penultimate layer of VGG19 (baseline). For the pruned layer we also report the average number of nodes selected (±SD across folds).

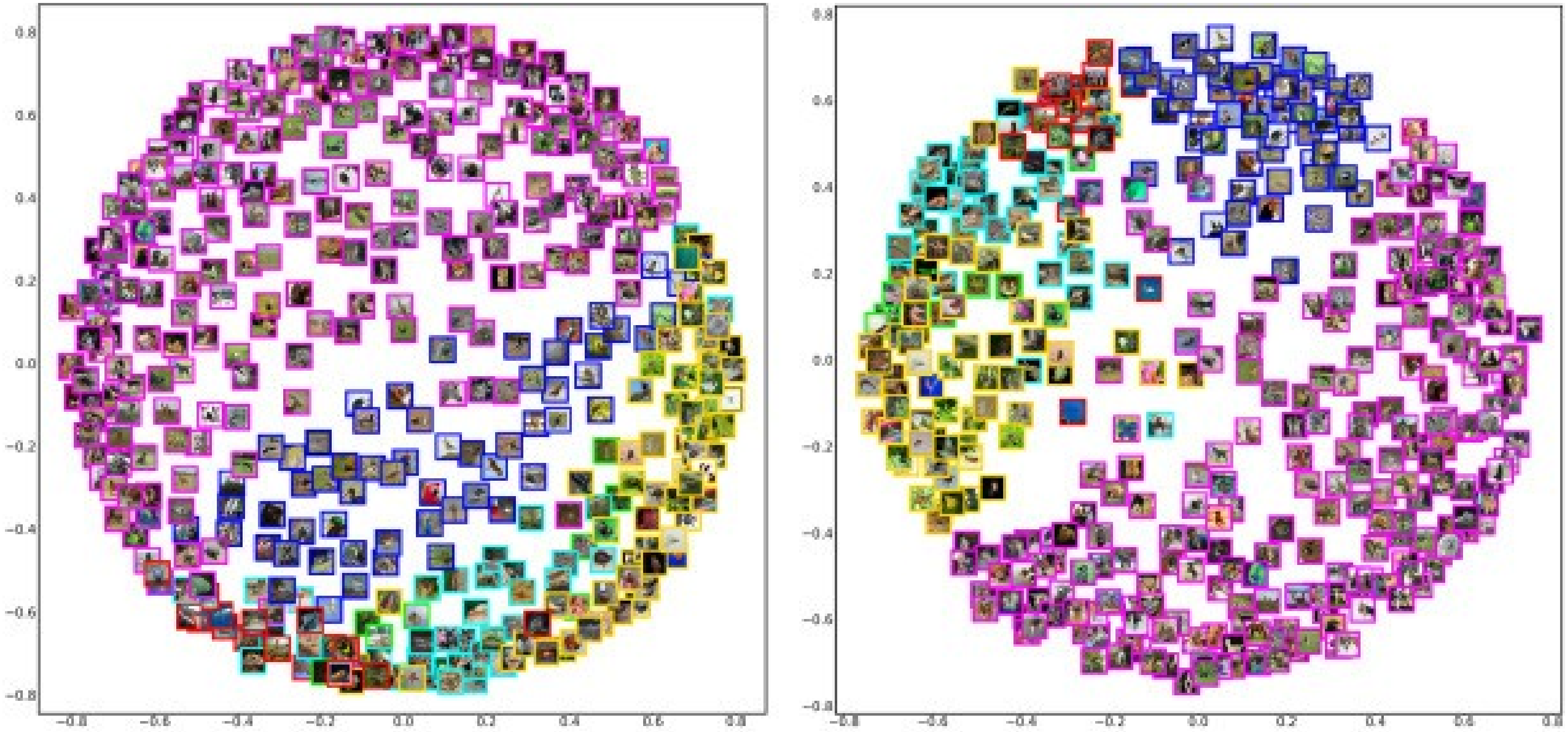
	Animals	Automobiles	Fruits	Furniture	Various	Vegetables
Baseline	0.61 (0.07)	0.51 (0.07)	0.33 (0.08)	0.29 (0.05)	0.43 (0.10)	0.32 (0.07)
PAG18	0.71 (0.09)	0.50 (0.05)	0.25 (0.15)	0.34 (0.08)	0.50 (0.13)	0.27 (0.07)
LASSO	0.64 (0.12)	0.51 (0.08)	0.38 (0.13)	0.37 (0.11)	0.47 (0.12)	0.31 (0.08)
Sim-DR	0.64	**0.57**	0.30	0.33	0.50	0.30
Pruned	**0.75** (0.05)	0.55 (0.08)	**0.39** (0.08)	**0.38** (0.07)	**0.56** (0.1)	**0.41** (0.05)
# nodes	807 (63)	647 (45)	563 (76)	557 (101)	830 (44)	605 (190)

Improves representational space for out of sample image embeddings

We use an independent dataset of Animal images and apply MDS, color coding indicates main animal types (our post-hoc annotation) [Left]

We repeat but using only feature indices retained from pruning against the experimental Animals dataset. [Right]

Improved prediction of behavioral and neural similarity spaces using pruned DNNs.
Tarigopula P, Fairhall SL, Bavaresco A, Truong N, Hasson U.
Neural Networks, 168:89-104,



(a) Original VGG-19

(b) Pruned VGG-19

Multidimensional scaling plots of the embeddings corresponding to the 398 animal classes of ImageNet with Original VGG-19 and the same network pruned for Animals. Magenta-mammals, Yellow-invertebrates, Cyan-reptiles, Green-amphibian, Blue-bird, Red-fish.

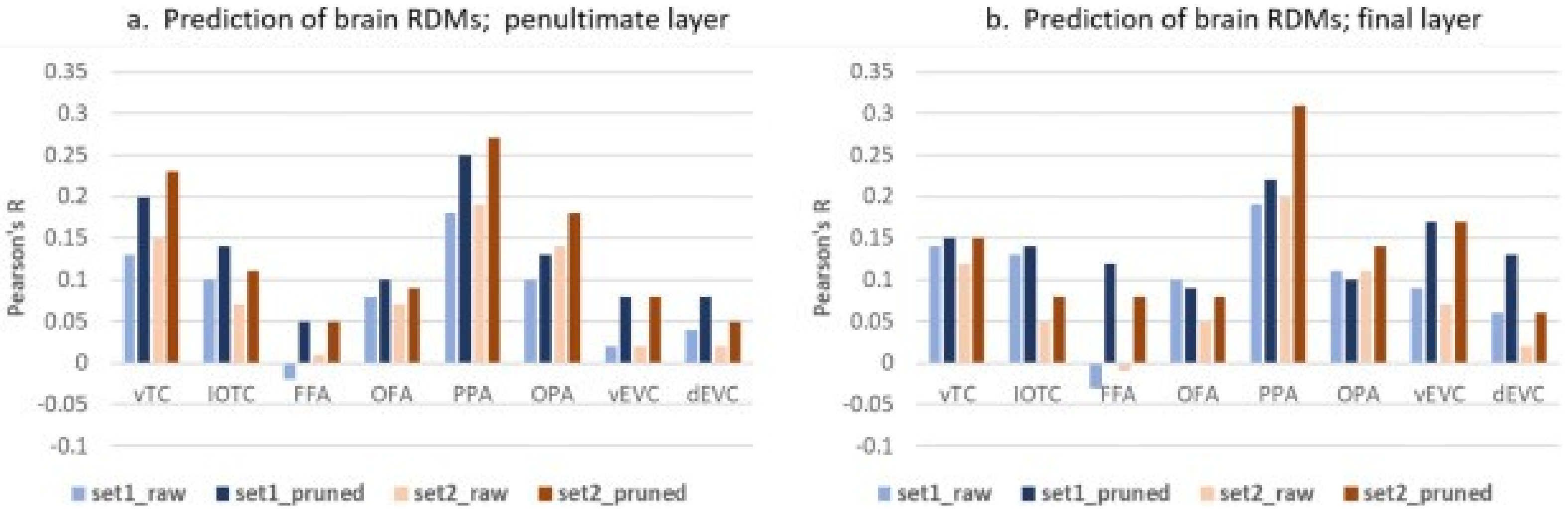
Improved representational space for out of sample brain data

Dataset from King et al. (fMRI)

- Two independent sets of 144 images.
- Produce RDMs from brain activity, per regions.
  Use RDMs to supervise pruning.
- Test on out of sample data.

Improved prediction of brain-derived RDMs, including FFA where full model fails.

```
Improved prediction of behavioral
and neural similarity spaces using
pruned DNNs.
Tarigopula P, Fairhall SL, Bavaresco
A, Truong N, Hasson U.
```
Neural Networks, 168:89-104, 2023



Learning pruning across image sets. Pruning was learned for DNN embeddings for one image set and applied to predict brain activity patterns associated with a different image set. (a) Prediction from embeddings from VGG-19 penultimate layer (b) Predictions from embeddings from VGG-19 final (1000-node) layer.

Are different prunings selecting for different information in the model?

Of the 4096 features analyzed different pruning retains less than 25%.

Question: are these feature sets coding for different information?

SOTA: Not necessarily, as DNNs are both *sparse* and *redundant.* Many nodes may not code for information at all, and remaining nodes may be correlated.

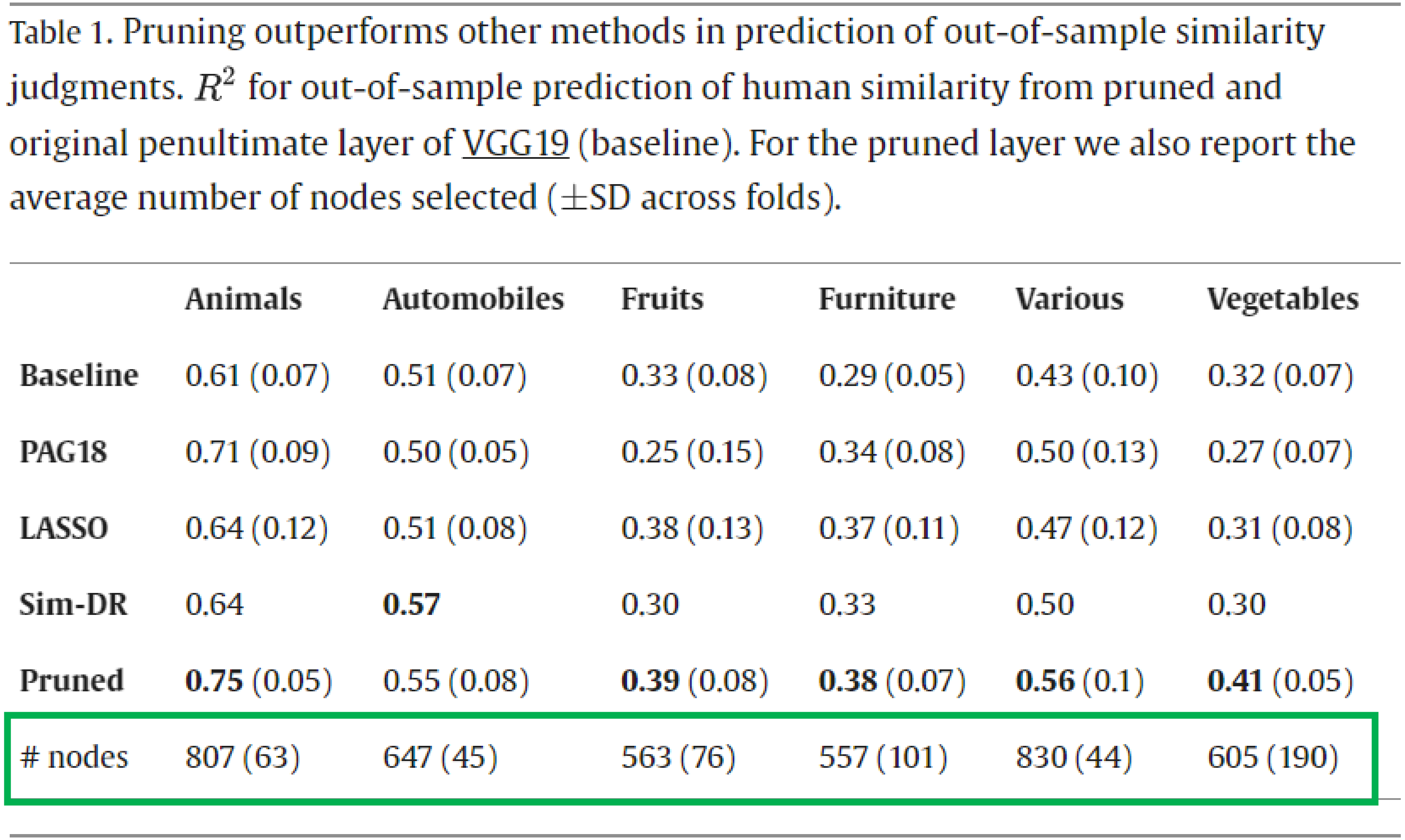


Table 1. Pruning outperforms other methods in prediction of out-of-sample similarity judgments. R^2 for out-of-sample prediction of human similarity from pruned and original penultimate layer of VGG19 (baseline). For the pruned layer we also report the average number of nodes selected (±SD across folds).

	Animals	Automobiles	Fruits	Furniture	Various	Vegetables
Baseline	0.61 (0.07)	0.51 (0.07)	0.33 (0.08)	0.29 (0.05)	0.43 (0.10)	0.32 (0.07)
PAG18	0.71 (0.09)	0.50 (0.05)	0.25 (0.15)	0.34 (0.08)	0.50 (0.13)	0.27 (0.07)
LASSO	0.64 (0.12)	0.51 (0.08)	0.38 (0.13)	0.37 (0.11)	0.47 (0.12)	0.31 (0.08)
Sim-DR	0.64	**0.57**	0.30	0.33	0.50	0.30
Pruned	**0.75** (0.05)	0.55 (0.08)	**0.39** (0.08)	**0.38** (0.07)	**0.56** (0.1)	**0.41** (0.05)
# nodes	807 (63)	647 (45)	563 (76)	557 (101)	830 (44)	605 (190)

Bavaresco A, Truong N, Hasson U.
In preparation

Different prunings select for different features

We treat the features identified via pruning as 6 different sets, and consider their indices.

The overlap between the (indices of the) features retained from the different category is often low, as determined by Dice Coefficient

DICE = 2 * (intersection/union)).

Lowest Dice(Fruits, Automobiles) = 0.13

Highest Dice(Fruits, Vegetables) = 0.28.

Bavaresco A, Truong N, Hasson U.
In preparation

Different prunings select for different information: information maximization analysis

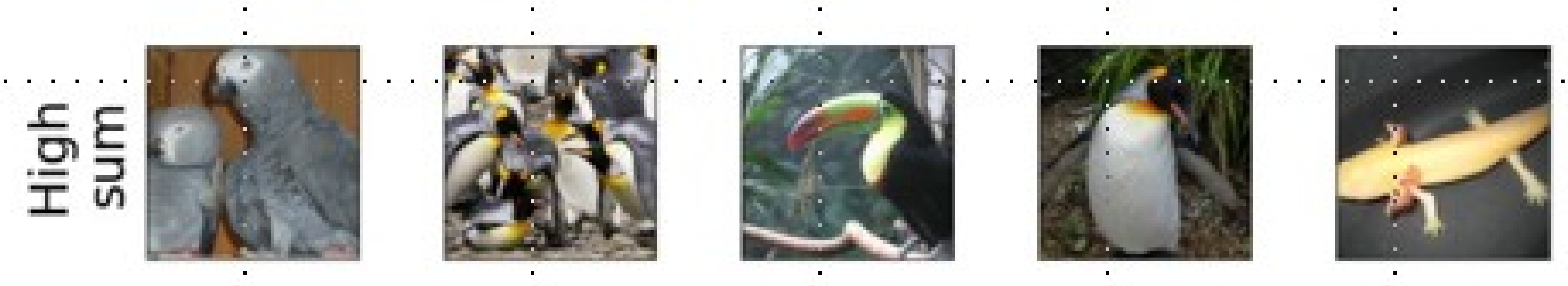
When used for feature extraction, **the selected features identify different information**.

We used a different 50,000 image dataset. Images were numerically represented using only activations on the retained features (6 different versions; 50Kx807; 50Kx647, etc’).

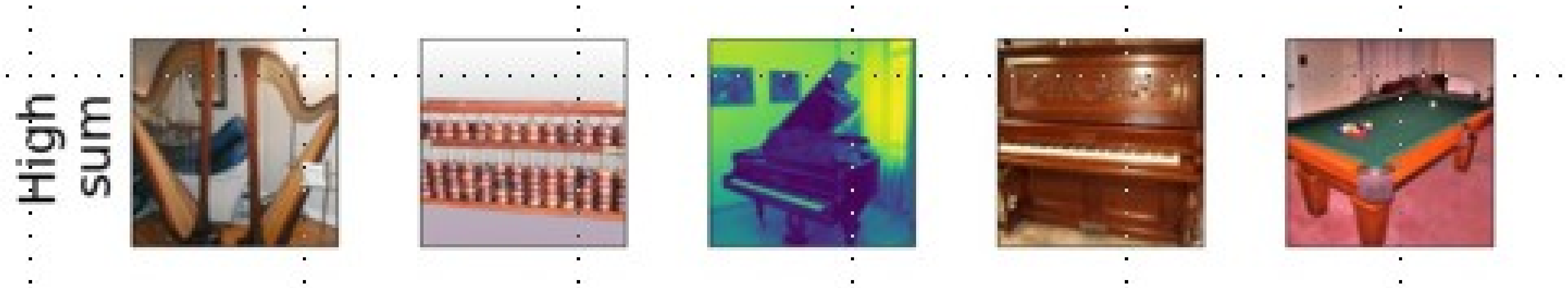
A test for information selectivity: For each of these derived embeddings we identified the top-5 images (of the 50K) that maximize activation on that set of features.

Bavaresco A, Truong N, Hasson U.
In preparation

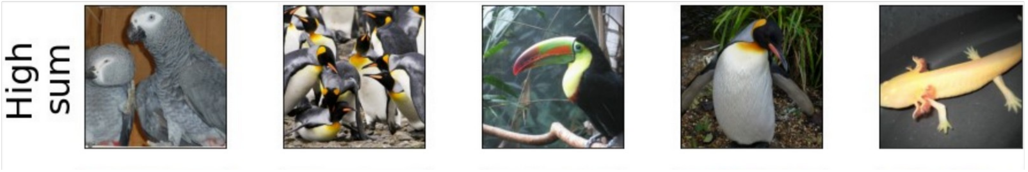
Top-5 images maximizing activations for features selected for animals (independent dataset)



Top-5 images maximizing activations for features selected for furniture (independent dataset)



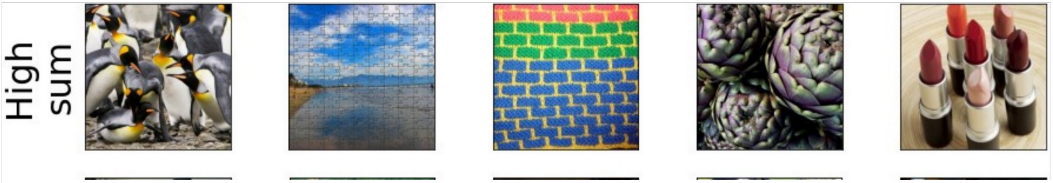
Animals



Furniture

Transportation

Various



Fruits

Vegetables

Finding: pruning-retained features, per category, are maximized by images that exemplify the category. Consistent with prototype theory

Retained features extract different latent dimensions

- We used the 50,000-image dataset. Images were represented only using the retained features (6 different versions).
- We applied PCA to each version and obtained the scores for the 50K images on the first Principal Component (PC1).
- We correlated the 50K PC1 scores across solutions.

Most correlations were medium-low.

Lowest Correlation(Transportation, Various) = 0.13.

Though some were high

Highest Correlation(Fruits, Vegetables) = 0.90

Conclusion: Different retained sets select for different latent dimensions

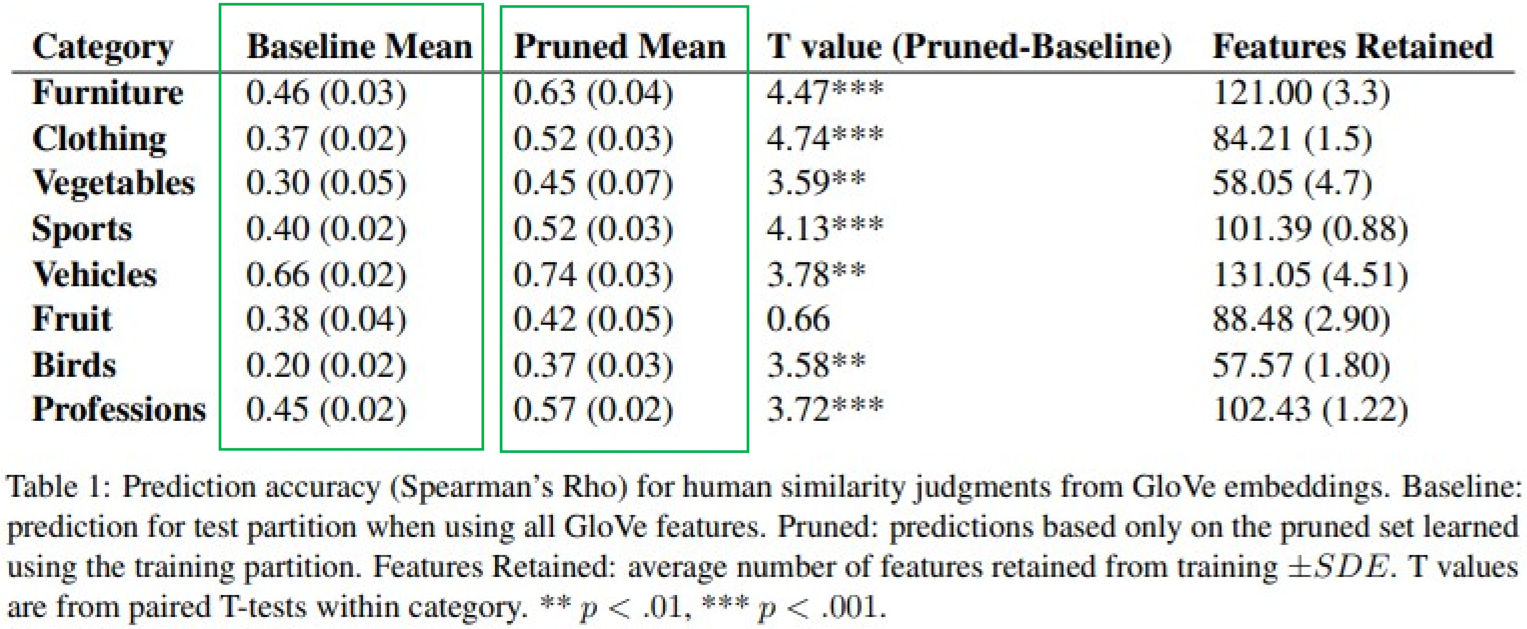
Bavaresco A, Truong N, Hasson U.
In preparation

Improved prediction of word-similarity judgments

Design: 8 noun categories each containing 20-30 words and human similarity judgments for all word pairs.

ML: Predict test-set using (1) all GloVe features, or (2) GloVe feature-sets pruned from training set.

Result: Pruning improves out of sample prediction using a less than half of GloVe’s 300 features.



Category	Baseline Mean	Pruned Mean	T value (Pruned-Baseline)	Features Retained
Furniture	0.46 (0.03)	0.63 (0.04)	4.47***	121.00 (3.3)
Clothing	0.37 (0.02)	0.52 (0.03)	4.74***	84.21 (1.5)
Vegetables	0.30 (0.05)	0.45 (0.07)	3.59**	58.05 (4.7)
Sports	0.40 (0.02)	0.52 (0.03)	4.13***	101.39 (0.88)
Vehicles	0.66 (0.02)	0.74 (0.03)	3.78**	131.05 (4.51)
Fruit	0.38 (0.04)	0.42 (0.05)	0.66	88.48 (2.90)
Birds	0.20 (0.02)	0.37 (0.03)	3.58**	57.57 (1.80)
Professions	0.45 (0.02)	0.57 (0.02)	3.72***	102.43 (1.22)

Table 1: Prediction accuracy (Spearman’s Rho) for human similarity judgments from GloVe embeddings. Baseline: prediction for test partition when using all GloVe features. Pruned: predictions based only on the pruned set learned using the training partition. Features Retained: average number of features retained from training $\pm SDE$. T values are from paired T-tests within category. ** $p < .01$, *** $p < .001$.

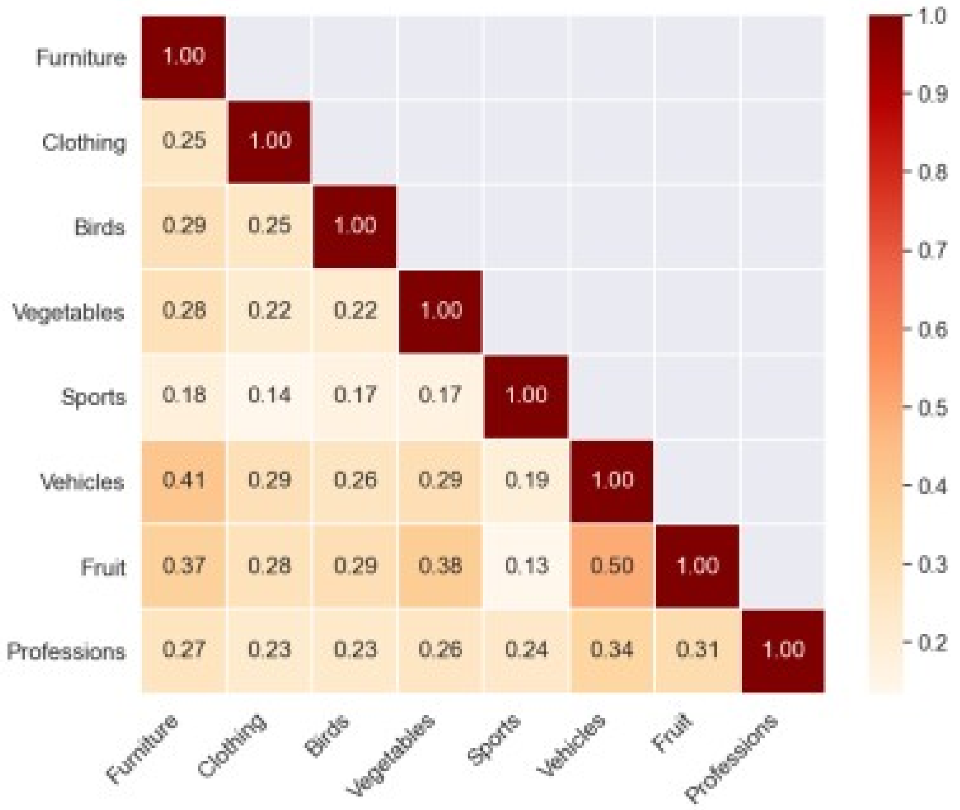
Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot, and Uri Hasson. 2023. Enhancing Interpretability Using Human Similarity Judgements to Prune Word Embeddings. In Proceedings of the 6th BlackboxNLP Workshop.

Improved prediction selects for different GloVe feature-sets

For each solution we identify the retained feature indices and compute Dice Coefficient.

Concordance (Dice) between pruned sets is often low.

Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot, and Uri Hasson. 2023. Enhancing Interpretability Using Human Similarity Judgements to Prune Word Embeddings. In Proceedings of the 6th BlackboxNLP Workshop.



	Furniture	Clothing	Birds	Vegetables	Sports	Vehicles	Fruit	Professions
Furniture	1.00							
Clothing	0.25	1.00						
Birds	0.29	0.25	1.00					
Vegetables	0.28	0.22	0.22	1.00				
Sports	0.18	0.14	0.17	0.17	1.00			
Vehicles	0.41	0.29	0.26	0.29	0.19	1.00		
Fruit	0.37	0.28	0.29	0.38	0.13	0.50	1.00	
Professions	0.27	0.23	0.23	0.26	0.24	0.34	0.31	1.00

Figure 1: Dice coefficient indicating overlap of features sets pruned by different categories

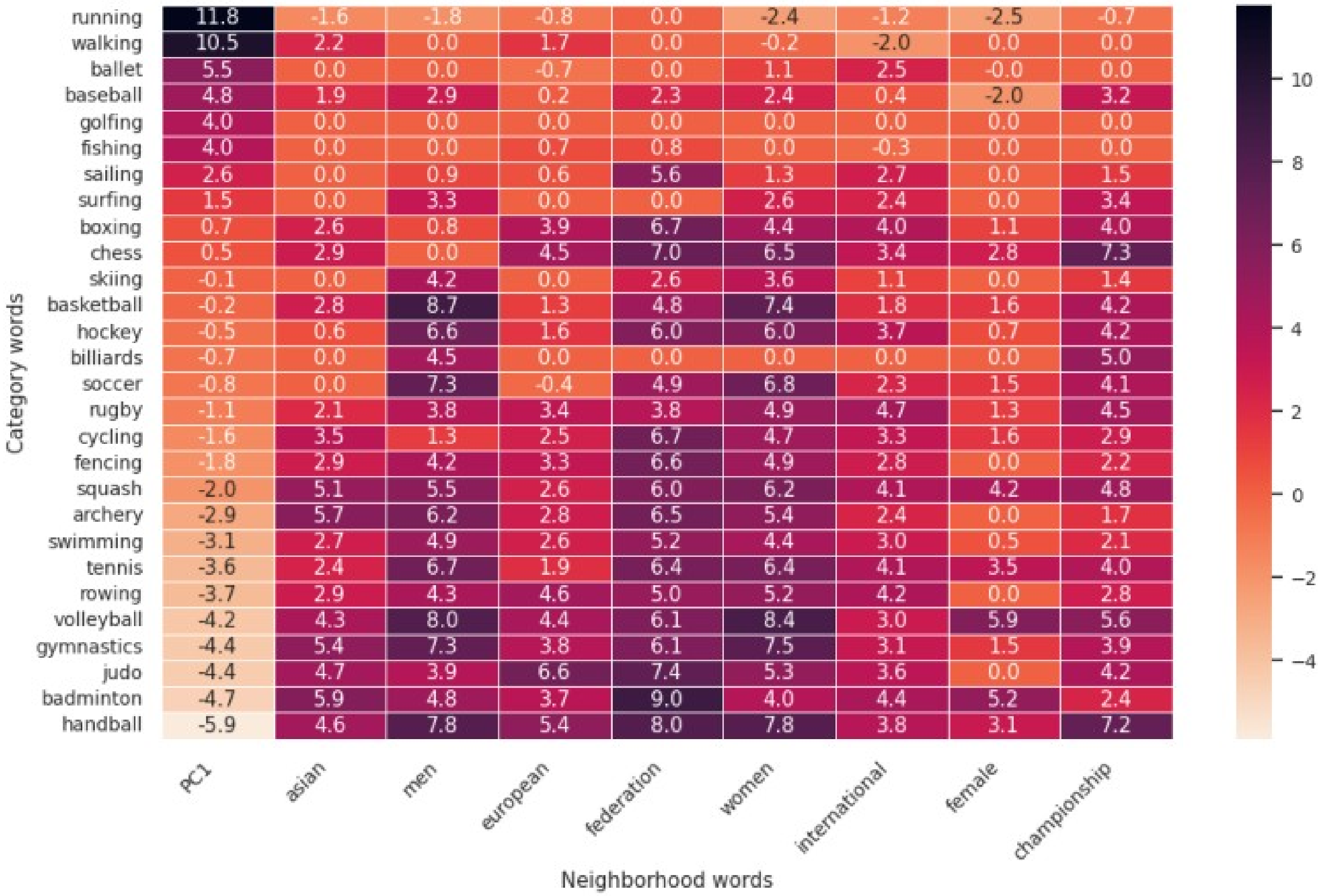
Interpretability of a retained feature-set: Example from SPORTS category

What semantics are captured by the retained embeddings.

We extract first PC from PCA analysis on retained embeddings (scores on column 1).

We find those corpus words whose co-occurrence pattern with the different sports correlates with these rankings (co-occurrence of adjectives).

Results suggest that human similarity judgments are sensitive to gender-, location-inclusiveness and (relatedly)



	PC1	asian	men	european	federation	women	international	female	championship
running	11.8	-1.6	-1.8	-0.8	0.0	-2.4	-1.2	-2.5	-0.7
walking	10.5	2.2	0.0	1.7	0.0	-0.2	-2.0	0.0	0.0
ballet	5.5	0.0	0.0	-0.7	0.0	1.1	2.5	-0.0	0.0
baseball	4.8	1.9	2.9	0.2	2.3	2.4	0.4	-2.0	3.2
golfing	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
fishing	4.0	0.0	0.0	0.7	0.8	0.0	-0.3	0.0	0.0
sailing	2.6	0.0	0.9	0.6	5.6	1.3	2.7	0.0	1.5
surfing	1.5	0.0	3.3	0.0	0.0	2.6	2.4	0.0	3.4
boxing	0.7	2.6	0.8	3.9	6.7	4.4	4.0	1.1	4.0
chess	0.5	2.9	0.0	4.5	7.0	6.5	3.4	2.8	7.3
skiing	-0.1	0.0	4.2	0.0	2.6	3.6	1.1	0.0	1.4
basketball	-0.2	2.8	8.7	1.3	4.8	7.4	1.8	1.6	4.2
hockey	-0.5	0.6	6.6	1.6	6.0	6.0	3.7	0.7	4.2
billiards	-0.7	0.0	4.5	0.0	0.0	0.0	0.0	0.0	5.0
soccer	-0.8	0.0	7.3	-0.4	4.9	6.8	2.3	1.5	4.1
rugby	-1.1	2.1	3.8	3.4	3.8	4.9	4.7	1.3	4.5
cycling	-1.6	3.5	1.3	2.5	6.7	4.7	3.3	1.6	2.9
fencing	-1.8	2.9	4.2	3.3	6.6	4.9	2.8	0.0	2.2
squash	-2.0	5.1	5.5	2.6	6.0	6.2	4.1	4.2	4.8
archery	-2.9	5.7	6.2	2.8	6.5	5.4	2.4	0.0	1.7
swimming	-3.1	2.7	4.9	2.6	5.2	4.4	3.0	0.5	2.1
tennis	-3.6	2.4	6.7	1.9	6.4	6.4	4.1	3.5	4.0
rowing	-3.7	2.9	4.3	4.6	5.0	5.2	4.2	0.0	2.8
volleyball	-4.2	4.3	8.0	4.4	6.1	8.4	3.0	5.9	5.6
gymnastics	-4.4	5.4	7.3	3.8	6.1	7.5	3.1	1.5	3.9
judo	-4.4	4.7	3.9	6.6	7.4	5.3	3.6	0.0	4.2
badminton	-4.7	5.9	4.8	3.7	9.0	4.0	4.4	5.2	2.4
handball	-5.9	4.6	7.8	5.4	8.0	7.8	3.8	3.1	7.2

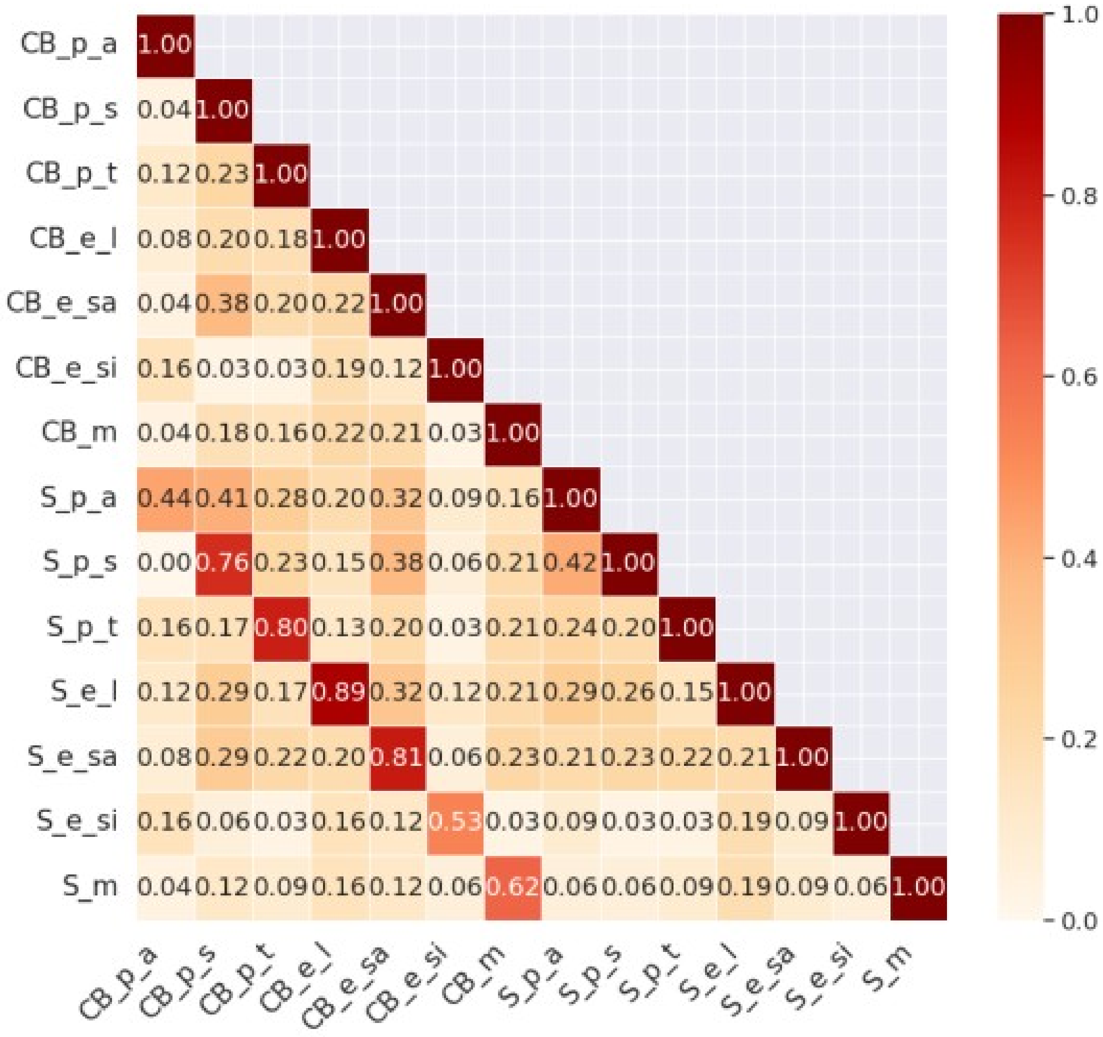
Figure 2: PMI values for words that correlate with co-hyponym scores on first PC computed from pruned embeddings

Pruning identifies different verb representations in blind and sighted.

Design: 7 verb categories each containing 14 verbs and human similarity judgments for all verb-pairs. Judgments made by **Congenitally Blind** and **Sighted**.
ML: Prune GloVe embeddings to improve out-of-sample prediction.

Result: For verbs describing emission of animate sounds (e.g., whine) and light (e.g., blink) good concordance between retained features for blind and sighted. For others, e.g., perception sight (e.g., see, look) a lower concordance

W. Bao & U. Hasson. (2024). Identifying and Interpreting Non-Aligned Human Conceptual Representations using Language Modeling. Proceedings of the ICLR 2024 Workshop Re-Align



	CB_p_a	CB_p_s	CB_p_t	CB_e_l	CB_e_sa	CB_e_si	CB_m	S_p_a	S_p_s	S_p_t	S_e_l	S_e_sa	S_e_si	S_m
CB_p_a	1.00													
CB_p_s	0.04	1.00												
CB_p_t	0.12	0.23	1.00											
CB_e_l	0.08	0.20	0.18	1.00										
CB_e_sa	0.04	0.38	0.20	0.22	1.00									
CB_e_si	0.16	0.03	0.03	0.19	0.12	1.00								
CB_m	0.04	0.18	0.16	0.22	0.21	0.03	1.00							
S_p_a	0.44	0.41	0.28	0.20	0.32	0.09	0.16	1.00						
S_p_s	0.00	0.76	0.23	0.15	0.38	0.06	0.21	0.42	1.00					
S_p_t	0.16	0.17	0.80	0.13	0.20	0.03	0.21	0.24	0.20	1.00				
S_e_l	0.12	0.29	0.17	0.89	0.32	0.12	0.21	0.29	0.26	0.15	1.00			
S_e_sa	0.08	0.29	0.22	0.20	0.81	0.06	0.23	0.21	0.23	0.22	0.21	1.00		
S_e_si	0.16	0.06	0.03	0.16	0.12	0.53	0.03	0.09	0.03	0.03	0.19	0.09	1.00	
S_m	0.04	0.12	0.09	0.16	0.12	0.06	0.62	0.06	0.06	0.09	0.19	0.09	0.06	1.00

Summary

Human representations of concepts/categories are better approximated by AI models that are modified to reflect only the relevant variance dimensions, as indicated by specific subsets of features.

Practically, pruning via sequential feature selection is one way to identify these dimensions.

- Competitive in learning human representations.
- Allows interpreting the relevant lower dimensions.