IMPROVING ISOMORPHISMS

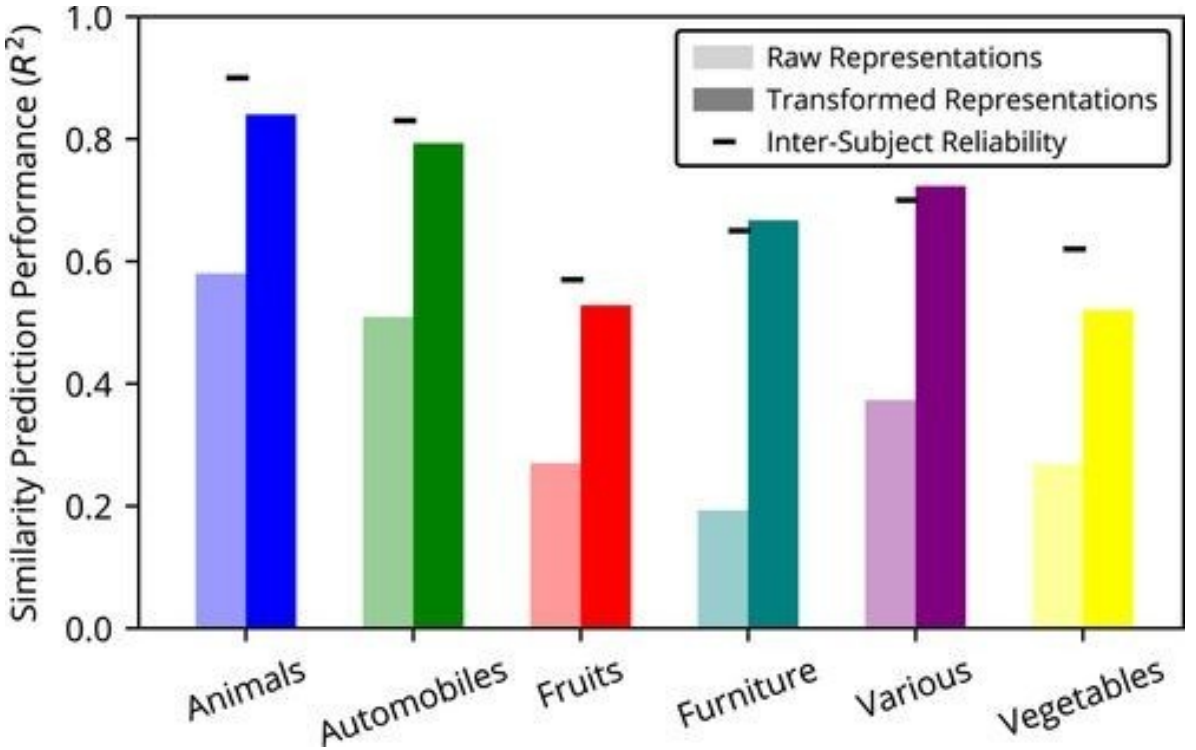
PRUNING VS. REWEIGHTING

OBJECTIVES

- Implementation details of supervised feature pruning vs. feature reweighting (Tarigopula et al.)
- Why pruning works? Prior work on sparsity and its relation to categorization (Hu et al)
- Sparsity and its relation to network representation (Truong and Hasson)

ISOMORPHISM BETWEEN DNNS AND HUMAN JUDGMENT IS LOW. WHY?

- Peterson: DNNs develop the correct basis set of features, just at the wrong level of saliency.
- Learning a reweighting of salience (transformed representation) improves prediction of human similarity judgments



REMINDER: EVALUATING THE CORRESPONDENCE BETWEEN REPRESENTATIONS

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

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

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

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

SIMILARITY MATRIX from humans

FEATURE MATRIX (dot product/ R/ cosine)

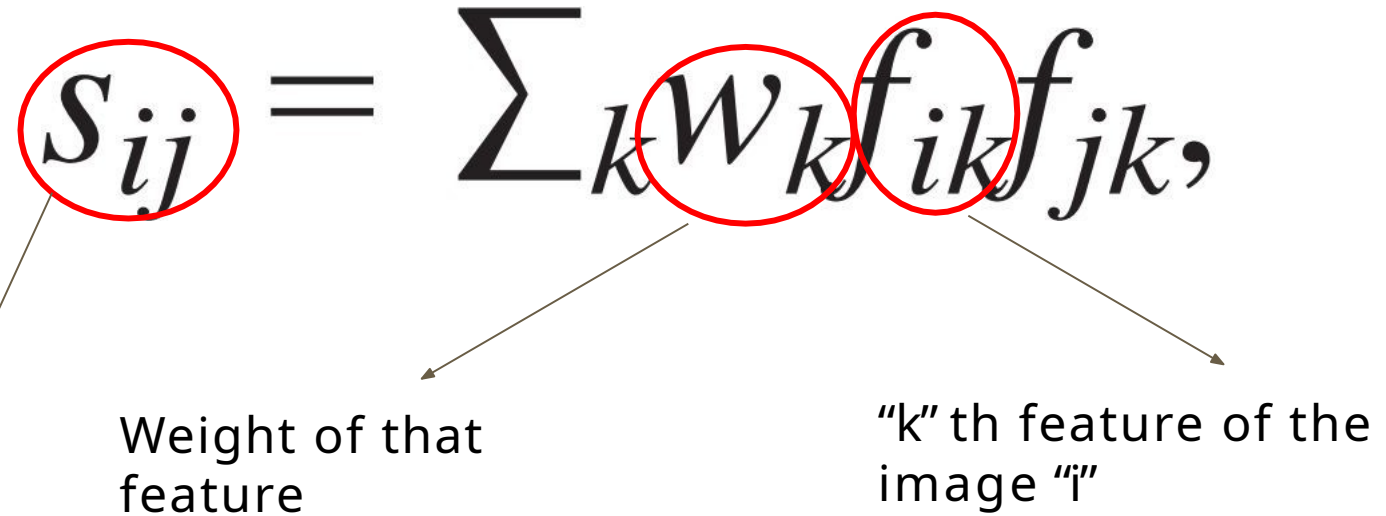
Reminder: Transforming Representations

- 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

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

Reminder: Transforming Representations

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



TRANSFORMATION ARE NOW FREQUENTLY USED

- Other forms of linear-transforms of the embedding matrix from images have been studied (Attarian et al., 2020).
- In computational linguistics, reweighting of word-embedding vectors were shown to improve prediction of human similarity judgments (Richie and Bhatia, 2020).
- Non-linear reweighting was also shown to be effective in modeling similarity judgments (Sanders and Nosofsky, 2020)

Attarian, M., Roads, B. D., and Mozer, M. C. (2020). Transforming neural network visual representations to predict human judgments of similarity.arXiv preprint arXiv:2010.06512.

Richie, R. and Bhatia, S. (2020). Similarity judgment within and across categories: A comprehensive model comparison

Sanders, C. A. and Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. Computational Brain & Behavior, pages 1–2

ARE TRANSFORMATIONS NECESSARY

- Reweighting operationalizes the assumption that DNNs learn relevant features but assign them different (i.e., wrong/mismatching) levels of salience with respect to humans.
- A different possibility: DNNs **do acquire** the relevant features at appropriate levels of salience.
  - It is just that that in any particular test-context where human similarity-space is predicted, the contribution of relevant features is diluted by irrelevant ones.
  - Consider predicting human similarity judgments of animals or tools (both RDMS) from brain activity (an RDM). Would you use the entire brain as feature set for constructing the RDM?
- Alternative idea: taking the entire penultimate layer of a DNN as the relevant basis set unintentionally combines two representational sub-spaces: those relevant *for a given set* of human similarity judgments and those less relevant

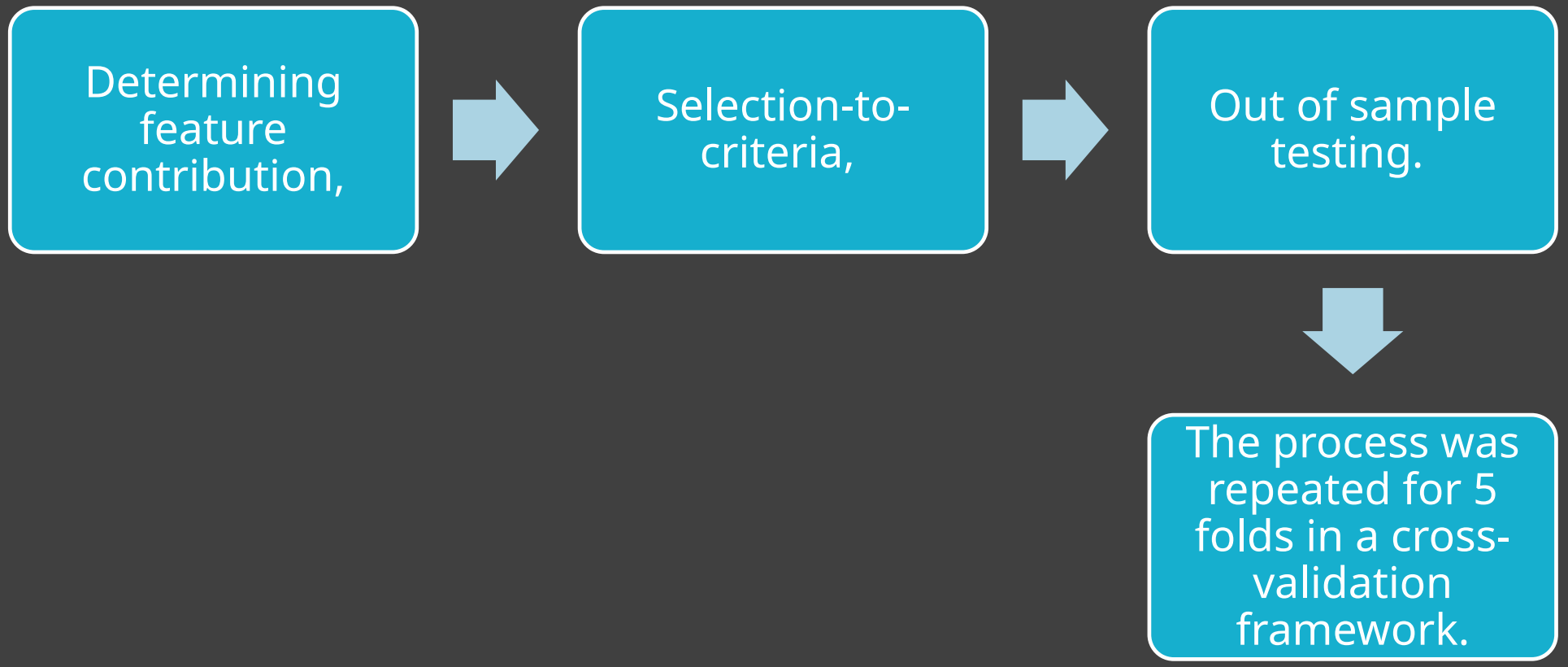
FEATURES AS BASIS SETS

Tarigopula, P., Fairhall, S. L., Bavaresco, A., Truong, N., & Hasson, U. (2023). Improved prediction of behavioral and neural similarity spaces using pruned DNNs. *Neural Networks*, *168*, 89-104

FORMALLY

- Hypothesis
  - Feature embeddings from the DNNs can account for human similarity spaces better than can be concluded from raw second-order-isomorphism.
  - Feature reduction (pruning) by ranking features can improve the fit between DNNs and similarity judgements
- General idea (non-technical intro)
  - Extract DNN image embeddings from the penultimate layer of VGG-19
  - Identify a subset of features that improves prediction of human similarity judgments (or brain similarity patterns).
  - Use the same 6 categories and images from Peterson et al.

METHOD: SEQUENTIAL FEATURE SELECTION IN CV



DETERMINING FEATURE CONTRIBUTION (TRAIN SET ONLY)

- In each cross-validation iteration, 20% (n=24) of the images are designated as a test set and 80% (n=96) as train set.
- Baseline-2OI was defined as the train-set's 2OI between the DNN representation and human similarity judgments for those 96 images.
  - We quantified each feature's contribution to baseline-2OI by removing only that feature and recomputing train-set-2OI. The feature was then reinserted and the next removed till the process was repeated for all 4096 features.
  - Consequently,
    - **`important' features are those whose removal produces a 2OI value below baseline-2O**I and
    - **`irrelevant' features are those whose removal produces a 2OI value above baseline-2OI**.
    - This produces a rank order of each feature's independent importance to baseline-2OI.

SELECTION TO CRITERIA (TRAIN SET ONLY)

After feature ranking, we consecutively insert features, according to their importance rank, into a candidate feature set.

Each time a feature is added to the candidate set, we construct an RDM from the features in the set and recompute 2OI against the train-set human similarity judgments.

We add all features exhaustively and then we identify set of features associated with the maximal value reached.

The set identified this way constitutes the **pruned network associated with a specific fold**.

VALIDATION (TEST SET; OUT OF SAMPLE PREDICTION)

- Once a pruned node-set is determined, we apply it to the left-out test set.
- To begin, we define *test-set's baseline 20I* which is the R^2 produced when considering all 4096 nodes.
- **Evaluating the pruning:**
  - The test-set images are inserted into the DNN, and we extract the activation values for the selected nodes *in the pruned layer*.
  - We then construct an RDM for the test-set images and report pruned-net-2OI as R^2.
  - We evaluate this value in relation to the test-set's baseline 20I which is the R^2 produced when considering all 4096 nodes rather than the pruned subset.

RESULTS

PREDICTION OF HUMAN SIMILARITY JUDGMENTS

- Baseline: 2OI between the DNN and human RDMs prior to any reweighting/pruning, averaged across the five out-of-sample data for the test folds
- PAG18: ridge regression as implemented by Peterson 2018, applied to the five out-of-sample folds defined in our process.
- Sim-DR: a reweighting approach developed Jha et al 2020, which optimizes a projection of DNN embeddings to a lower-dimensional that matches human similarity judgments
- LASSO is our own variation of the reweighting implemented by PAG18 but which uses LASSO-regularized regression that is further constrained to only positive weights.
- Pruned: pruning method introduced here

PREDICTION OF HUMAN SIMILARITY JUDGMENTS

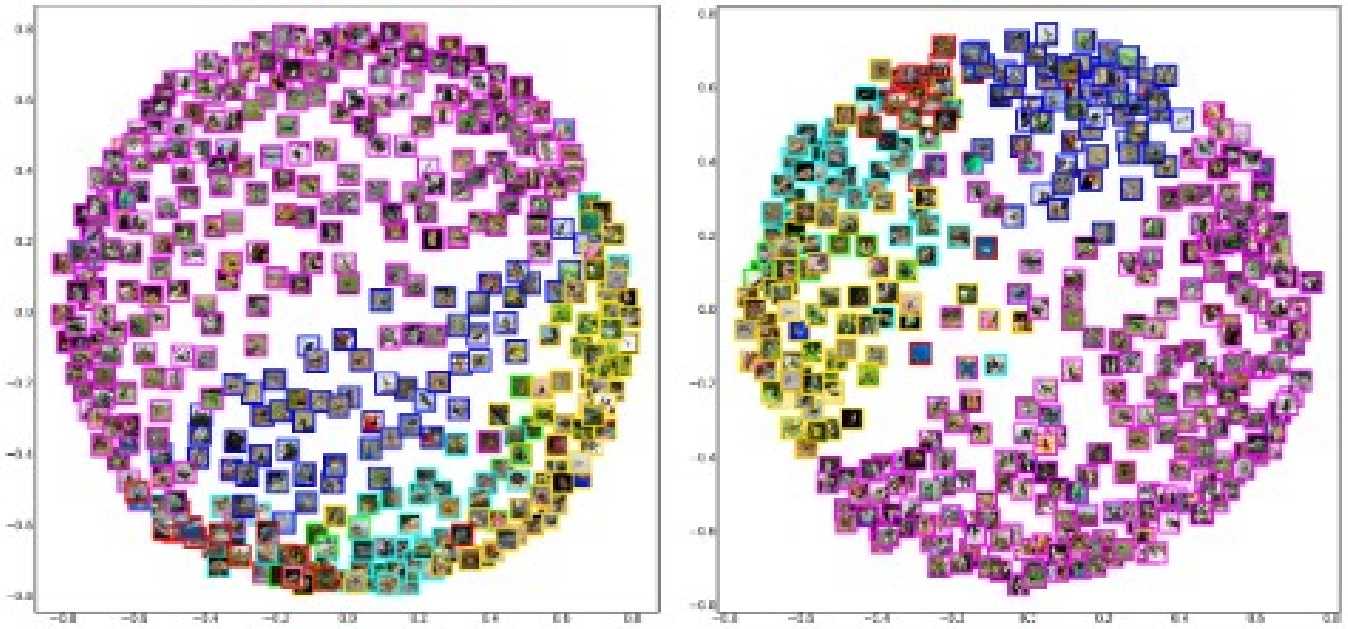
- Baseline: match between the DNN and human similarity space prior to any modification, averaged across the five out-of-sample data for the test folds
- PAG18: Reweighting. ridge regression as implemented by Peterson2018, applied to the five out-of sample folds used in our data.
- Sim-DR: a reweighting approach developed Jha et al 2020, which optimizes a projection of DNN embeddings to a lower-dimensional that matches human similarity judgments
- LASSO is our own variation of the reweighting implemented by PAG18 but which uses LASSO-regularized regression that is further constrained to only positive weights.
- Pruned: pruning method introduced here

Table 1: 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 ($\pm$ 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)

LOWER-DIMENSIONS OF HUMAN SIMILARITY JUDGMENTS

Multi Dimensional Scaling (MDS) plots of the non-pruned network and the network pruned for animals, on the 398 ANIMAL categories of ImageNet



(a) Original VGG-19 (b) Pruned VGG-19

Figure 1: Multi Dimensional 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

ORGANIZATION OF HUMAN SIMILARITY JUDGMENTS VS. WORDNET: WHAT IS WORDNET?

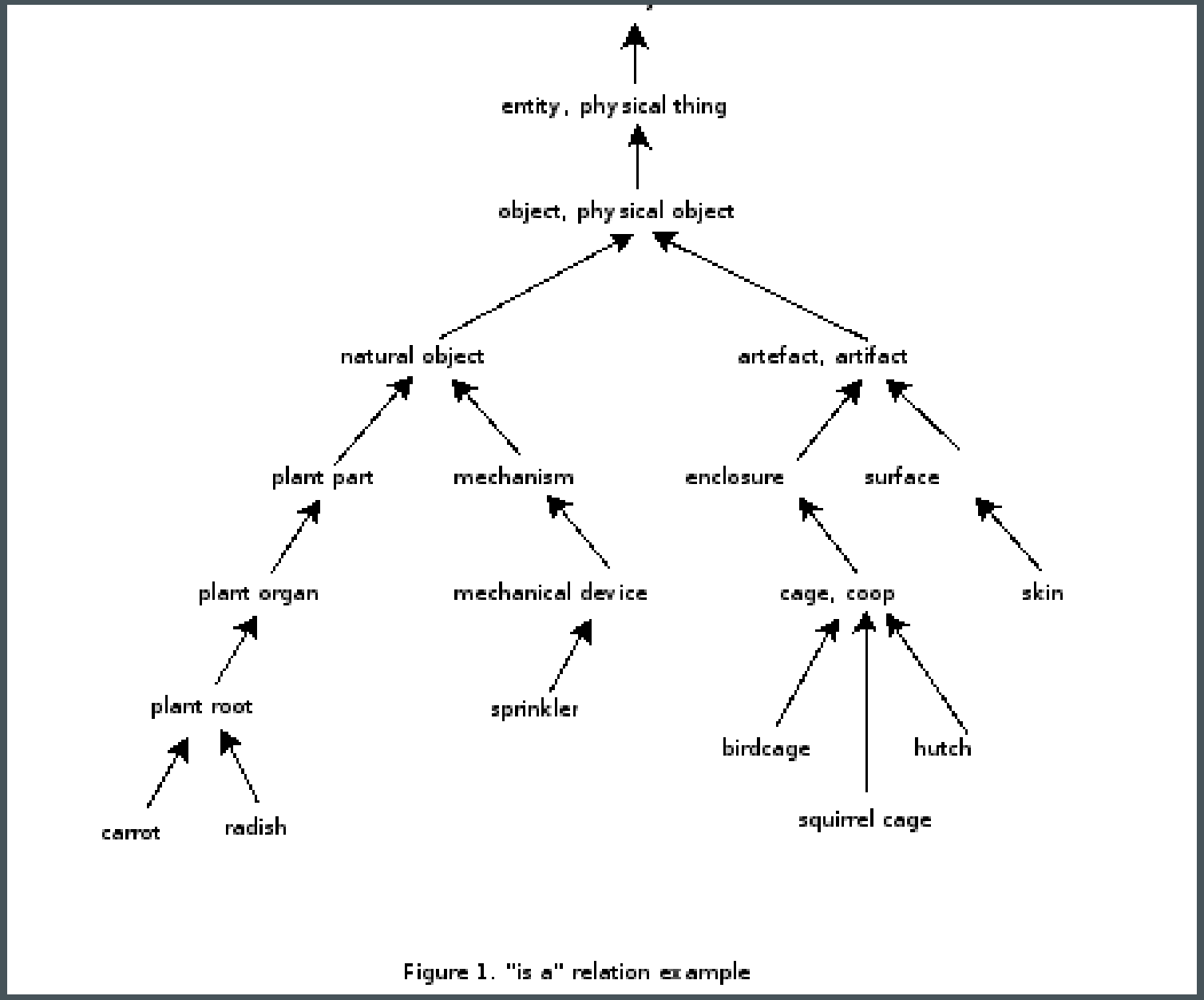


Figure 1. "is a" relation example

WORDNET RELATIONS

WE FOCUS ON HYPONYMY PATHS

Relation	Category	
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

ORGANIZATION OF HUMAN SIMILARITY JUDGMENTS VS. WORDNET

2OI reflects ranking of pair-wise similarities, but it does not directly address hierarchical structure.

Solution: we compare the hierarchical structure latent in the DNN similarity space to that of WordNet

ORGANIZATION OF HUMAN SIMILARITY JUDGMENTS VS. WORDNET. QUANTIFYING FIT

1

From DNN similarity space: produce clusters, and define ‘local neighborhood’ of each terminal leaf (image)

2

From Wordnet Graph: define ‘local neighborhood’ of each terminal leaf (image). Possible, because image-net categories match wordnet labels.

3

Compute: to what extent are neighbors in the DNN DAG also neighbors in the Wordnet Graph: Average ‘neighborhood fit’ across all nodes using Jaccard Index (set_intersection/ set_union)

ORGANIZATION OF HUMAN SIMILARITY JUDGMENTS VS. WORDNET

Table 2: Jaccard-Index concordance between a category's WordNet taxonomic neighborhood and its neighbors in DNN clusters. Higher values indicate greater agreement. Values shown for solutions across $N = 6 : 12$ DNN clusters. All comparisons statistically significant at $p < .01$ Bonferroni corrected for 6 comparisons.

	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$	$N = 11$	$N = 12$
Orig. Vgg-19	0.20	0.20	0.20	0.18	0.18	0.22	0.22
Pruned Vgg-19	0.30	0.26	0.26	0.25	0.26	0.26	0.26

'N' refers to number of clusters in DNN clustering solution (arbitrary)

HOW DO PRUNED AND REWEIGHTED NETS CLASSIFY?

TOP1/TOP5 OF P.A.G(18) RIDGE; LASSO (POS. WEIGHTS); PRUNING

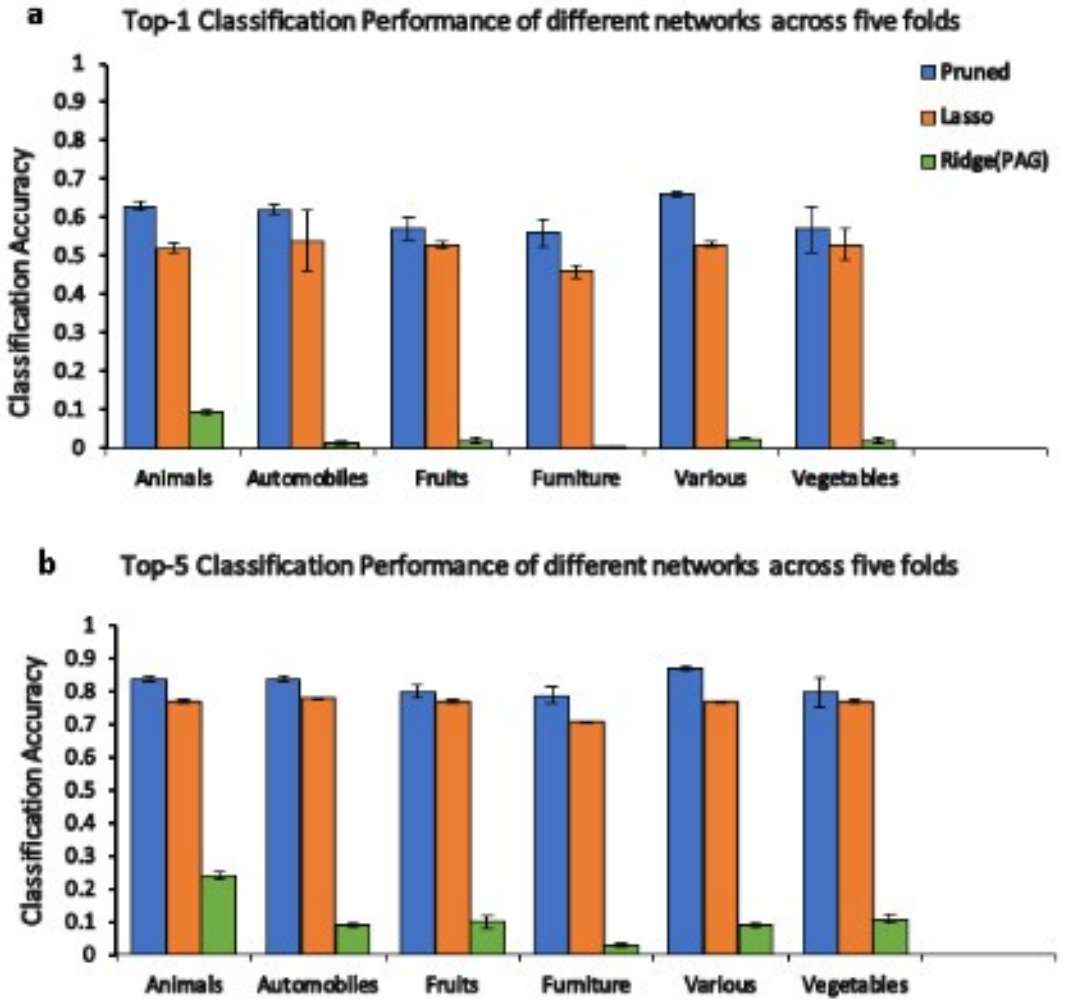


Figure 2: The classification performance of different networks across the five folds. (a) Top-1 performance (b) Top-5 performance

INTERIM CONCLUSIONS ON SUPERVISED PRUNING

- Can be used for (out of sample) prediction of human similarity spaces
- Outperformed previous reweighting-based methods for modifying network activation
- Better organized latent structure:
  - hierarchical clustering applied to pruned-embeddings proved a better fit to WordNet's hierarchy,
  - MDS spaces produced from pruned embeddings reflected a better clustering into basic-level semantic categories of Animals
  - Better match for Wordnet.
- Maintained top-1/top-5 classification accuracy at higher levels than regularized regression methods.
- Implications for theory:
  - DNNs already capture features relevant to human similarity judgments at an adequate level of salience, and for this reason, node activations do not need to be reweighted.
  - Need methods to filter out those features/nodes that are less relevant to modeling similarity of the domain at hand.

UNDERSTANDING BRAINS AND DNNS

PRUNING AS DIMENSIONALITY REDUCTION

INSIGHTS INTO BRAIN AND DNN ORGANIZATION

CHALLENGE: PRUNE DNNS TO BETTER EMULATE NEURAL SIMILARITY SPACES

RATHER THAN PRUNE DNNS FROM HUMAN SIMILARITY JUDGMENTS, PRUNE FROM ‘BRAIN SIMILARITY JUDGMENTS’

DATASET: FMRI ACTIVITY WHILE PEOPLED VIEWED 144 DIFFERENT IMAGES.

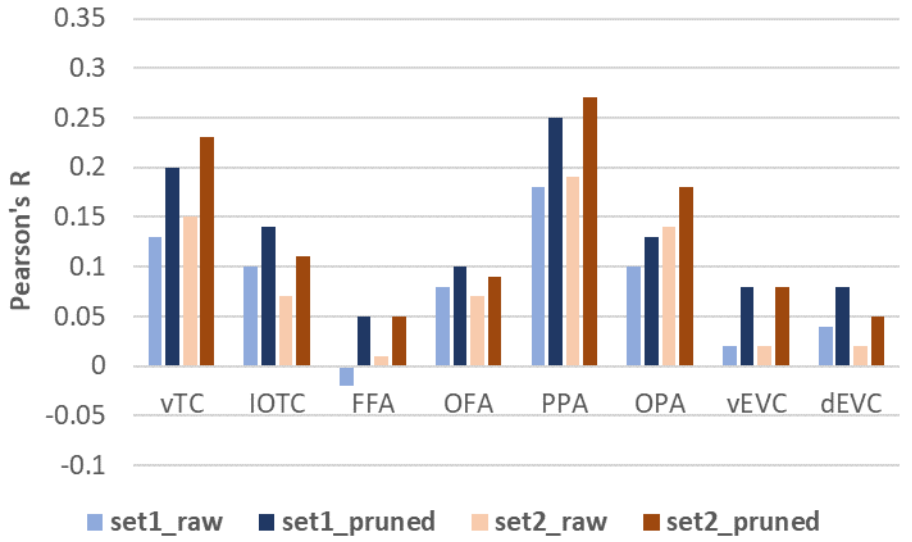
THE DATASET

Two sets of 144 images. Allows pruning based on one set and testing the performance of the pruning mask on the other (‘2 fold CV’)

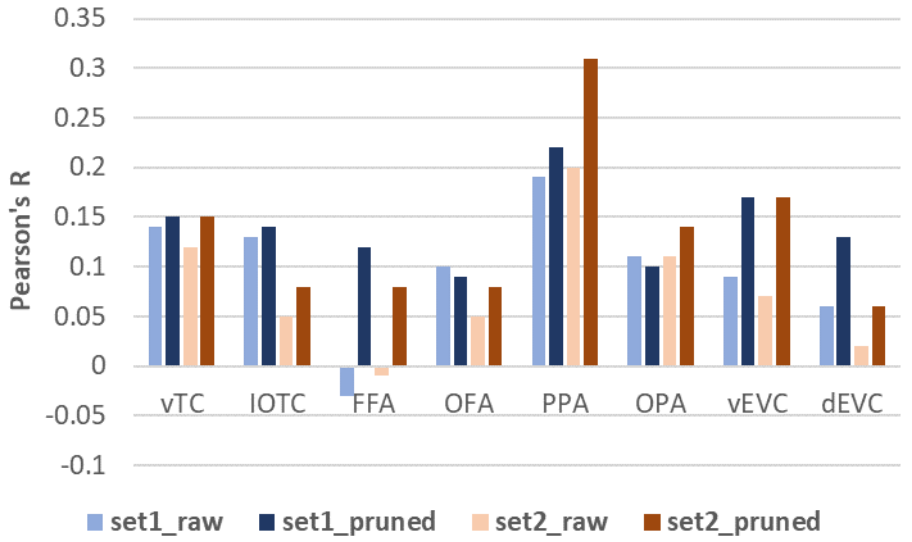
Data pre-collected and organized for 8 brain areas, some specialized for face and scene perception (FFA, PPA), and other spanning the visual cortex

PRUNING DNNS FROM ONE IMAGE SET, PREDICTING ON THE OTHER

a. Prediction of brain RDMs; penultimate layer



b. Prediction of brain RDMs; final layer



DIMENSIONALITY OF BRAIN ENCODING TRACKED BY NUMBER OF FEATURES RETAINED FROM DNN PRUNING

Table 4: Number of nodes retained from VGG-19 final layer for pruning based on different ROIs.

	vTC	lOTC	FFA	OFA	PPA	OPA	vEVC	dEVC
Set 1	27	29	10	60	75	40	41	71
Set 2	40	79	7	128	138	84	12	18

SUMMARY AND POINTS FOR THOUGHT

- Vision DNNs trained to classify already provide a moderate approximation of human representational space
- Reweighting is emerging as a new technique to improve this match (useful for practical applications), and was interpreted to suggest that DNNs learn human-like filters, but at wrong levels of salience
- Pruning outperforms reweighting in learning prediction of human representational spaces, but also originates in a different perspective on the importance of DNN filters: the filters are effective at the learned levels of salience, but different dataset benefit from different combinations of filters. Pruning is also more easily interpretable as a regularization (data reduction) technique in context of explainable AI and provides insights into brain organization.