Minimal Preprocessing Unifies Representational Similarity Rankings Across RSA, CKA, and CCA

Anonymous Authors

Editors:

Abstract

Similarity analyses of learned representations often produce different rankings across popular measures, which complicates comparison and reuse. We test whether a minimal and fully specified preprocessing step can reconcile these outcomes. Using three vision encoders and six text encoders on public datasets, we evaluate representation similarity analysis, linear centered kernel alignment, and singular vector canonical correlation analysis under raw features and under per feature z scoring across stimuli. On text encoders, standardization raises cross measure ranking agreement, for example Kendall tau between representation similarity analysis and centered kernel alignment increases from 0.64 to 0.89 for CLS pooled vectors, and between representation similarity analysis and canonical correlation analysis from 0.58 to 0.83. Mean pooled vectors already agree strongly and show smaller gains. In vision, heatmaps reveal that representation similarity analysis is sensitive to standardization while the other measures remain stable. A linear transfer probe on text shows positive associations between similarity and the negative of prediction error. An orthogonal transform control leaves centered kernel alignment unchanged, consistent with theory. These results support a simple reporting standard: state and apply dataset wise centering and variance scaling when comparing representations, since this improves agreement across measures and clarifies links to transfer.

Keywords: Representational similarity, RSA, CKA, CCA, standardization, z scoring, linear transfer, neural representations, geometry

1. Introduction

Systems that process related content often develop related geometry in their internal representations. This occurs in artificial networks and in neural data. Quantifying the relation between two learned representations is therefore a central tool. The literature offers several measures with different invariances and numerical ranges (e.g., RSA, SVCCA, CKA), and new approaches continue to emerge (Kriegeskorte et al., 2008; Raghu et al., 2017; Kornblith et al., 2019; Klabunde et al., 2025). In practice, conclusions about which models are most similar can change with the measure. We ask whether a trivial preprocessing step can unify conclusions across popular choices.

2. Background and definitions

Let $X \in \mathbb{R}^{n \times d_X}$ and $Y \in \mathbb{R}^{n \times d_Y}$ be features for the same n stimuli. We compare raw features to per-feature z-scoring across stimuli:

$$\tilde{X} = \left(X - \mathbf{1}\mu_X^\top\right) \operatorname{diag}(\boldsymbol{\sigma}_X)^{-1}, \qquad \tilde{Y} = \left(Y - \mathbf{1}\mu_Y^\top\right) \operatorname{diag}(\boldsymbol{\sigma}_Y)^{-1}, \tag{1}$$

where $(\mu, \boldsymbol{\sigma})$ denote columnwise mean and standard deviation.

RSA Build cosine RDMs D_X, D_Y, vectorize their strict upper triangles u_X, u_Y, and compute the Spearman correlation (Kriegeskorte et al., 2008):

$$\mathrm{RSA}(X, Y) = \rho_s(u_X, u_Y).$$

Linear CKA Normalized cross-covariance of centered features (Kornblith et al., 2019):

$$\mathrm{CKA}(X, Y) = \frac{\|X_c^\top Y_c\|_F^2}{\|X_c^\top X_c\|_F \, \|Y_c^\top Y_c\|_F}, \tag{2}$$

invariant to orthogonal transforms and uniform rescaling and closely related to centered RSA (Williams, 2024).

SVCCA Average canonical correlations after SVD reduction (Raghu et al., 2017).

Agreement and transfer Metric agreement uses Kendall's τ:

$$\tau = \frac{\#\text{concordant} - \#\text{discordant}}{\binom{M}{2}}. \tag{3}$$

For "what-for," fit a ridge map on train data and evaluate on test:

$$W^\star = \arg\min_W \|X_{\mathrm{tr}} W - Y_{\mathrm{tr}}\|_F^2 + \alpha \|W\|_F^2, \tag{4}$$

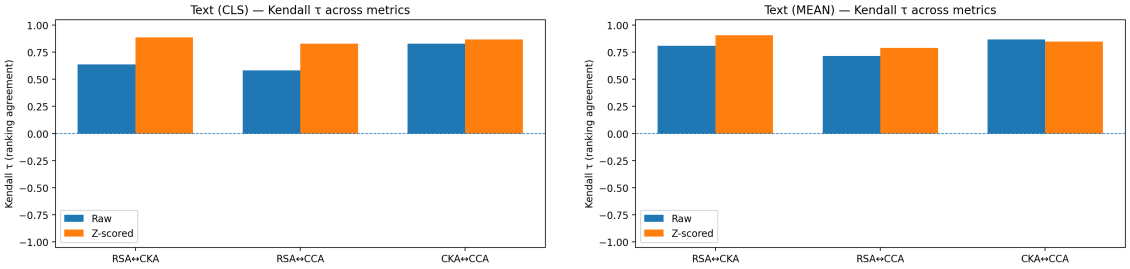
then report Spearman correlation between similarity and the negative of normalized MSE (Harvey et al., 2024).

3. Experimental setup

We use an excerpt of AG News for text and the test split of CIFAR ten for vision (Zhang et al., 2015; Krizhevsky, 2009). Each set contains about one thousand stimuli. Text encoders include DistilBERT Base Uncased, ALBERT Base v2, ELECTRA Small Discriminator, SqueezeBERT Uncased, all MiniLM L6 v2, and a very small BERT variant. Vision encoders include ResNet 50 and two CLIP ViT models that use B 16 and B 32 patching. For text we extract the last hidden layer and form CLS pooled and mean pooled sentence vectors. For vision we use the penultimate or global vector that precedes the final classifier. We compare raw features with the standardized features in equation (1). We compute representation similarity analysis with cosine based dissimilarity and Spearman correlation, linear centered kernel alignment as in equation (2), and singular vector canonical correlation analysis with a low rank of sixty four. For each family we obtain similarities for all model pairs and then compare the induced rankings through Kendall tau in equation (3). For the transfer analysis we use the ridge map in equation (4) and report the association between similarity and negative error.

4. Results

On text encoders the ranking agreement across measures increases after standardization. For CLS pooled vectors the Kendall coefficient between representation similarity analysis and centered kernel alignment rises from 0.64 to 0.89, and the coefficient between representation similarity analysis and canonical correlation analysis rises from 0.58 to 0.83. The agreement between centered kernel alignment and canonical correlation analysis is high in



(*a*) Text CLS: Kendall τ for raw vs. standardized.

(*b*) Text mean pooled: Kendall τ for raw vs. standardized.

Figure 1: Agreement of rankings across RSA, CKA, and CCA.

Table 1: Kendall tau for ranking agreement among measures for text encoders.

Case	RSA with CKA	RSA with CCA	CKA with CCA
CLS raw	0.64	0.58	0.83
CLS z	0.89	0.83	0.87
Mean raw	0.81	0.71	0.87
Mean z	0.90	0.79	0.85

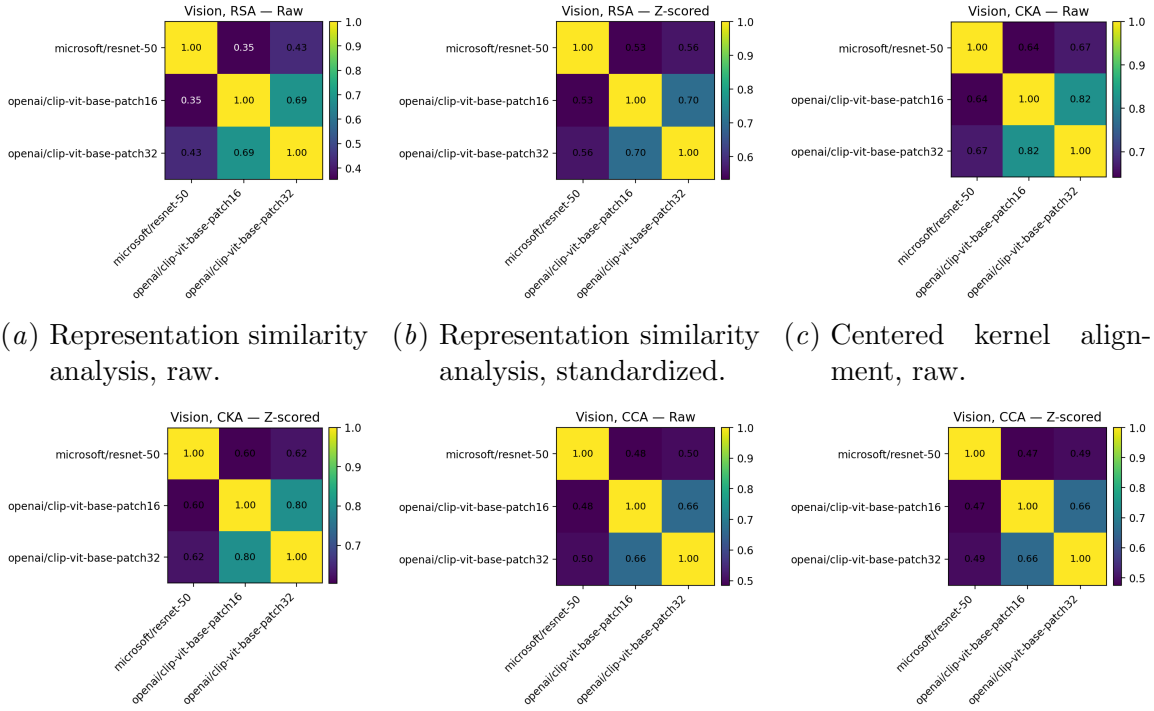
both cases. For mean pooled vectors the agreement is high in the raw case and improves modestly after standardization. Figure 1 summarizes both cases.

On vision encoders representation similarity analysis is sensitive to standardization while centered kernel alignment and canonical correlation analysis remain stable. Heatmaps in Figure 2 show that several off diagonal entries for representation similarity analysis increase after standardization, for example the pair of ResNet 50 and CLIP ViT B 16 moves from about 0.35 to about 0.53. With three models there are only three pairs and the Kendall coefficient among measures is one in both the raw and standardized cases.

Table 1 lists the agreement numbers for text. Table 2 shows that on text, higher similarity values are associated with better linear transfer as measured by rank correlation with the negative error. This supports the idea that standardized similarity is informative for reuse and stitching tasks. We also apply an orthogonal transform to one representation and observe that centered kernel alignment remains exactly the same while representation similarity analysis is numerically unchanged within tolerance, which matches known invariances.

5. Discussion

The results indicate that a small and clearly stated preprocessing step can reconcile conclusions across common measures. Per feature z scoring across stimuli removes nuisance offsets and scale differences and therefore reduces the ways in which measures can disagree. The effect is largest for CLS pooled text vectors and is visible in vision through the change



(*a*) Representation similarity analysis, raw.

(*b*) Representation similarity analysis, standardized.

(*c*) Centered kernel alignment, raw.

(*d*) Centered kernel alignment, standardized.

(*e*) Canonical correlation analysis, raw.

(*f*) Canonical correlation analysis, standardized.

Figure 2: Vision pairwise similarity heatmaps for ResNet 50, CLIP ViT-B16, and B32.

Table 2: Association between similarity and negative transfer error on text, mean pooled and standardized.

Measure	RSA	CKA	CCA
Spearman rho	0.54	0.46	0.35

in representation similarity analysis heatmaps. The positive association between similarity and transfer on text connects the measurement side with a practical goal. This observation is consistent with a decoding view which relates similarity to the alignment of linear readouts (Harvey et al., 2024), and is in line with evidence that metrics capturing overall representational geometry align more closely with differences in model behavior (Bo et al., 2024). The main limitation is the small number of vision models which forces a coarse agreement grid. Future work should enlarge the panel of models or treat layers as items to obtain a richer ranking structure. It will also be useful to evaluate kernel versions of centered kernel alignment and canonical correlation analysis and to extend the transfer analysis to matched tasks in vision.

Reproducibility statement

All datasets and models are public through the Hugging Face ecosystem. We use AG News and CIFAR ten and common encoders listed in Section 3 (Zhang et al., 2015; Krizhevsky, 2009). A single Colab notebook produces all figures and tables shown here, including feature extraction, the three similarity computations, ranking agreement, and the transfer probe. The code base will be released upon acceptance. Until then it is available from the authors upon reasonable request.

References

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008. URL: https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008/full.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97, 2019. URL: https://proceedings.mlr.press/v97/kornblith19a.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv:1706.05806*, 2017. URL: https://arxiv.org/abs/1706.05806.

Alexander H. Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlation analysis. In *Proceedings of the UniReps Workshop at NeurIPS 2024*, 2024. URL: https://openreview.net/forum?id=zMdnnFasgC.

Sarah E. Harvey, David Lipshutz, and Alexander H. Williams. What representational similarity measures imply about decodable information. *arXiv:2411.08197*, 2024. URL: https://arxiv.org/abs/2411.08197.

Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv:2411.14633*, 2024. URL: https://arxiv.org/abs/2411.14633.

Nathan Cloos, Guangyu Robert Yang, and Christopher J. Cueva. A framework for standardizing similarity measures in a rapidly evolving field. *arXiv:2409.18333*, 2024. URL: https://arxiv.org/abs/2409.18333.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus H. Maier-Hein, Markus Strohmaier, and Florian Lemmerich. ReSi: A comprehensive benchmark for representational similarity measures. *arXiv:2408.00531*, 2025. URL: https://arxiv.org/abs/2408.00531.

UniReps Organizers. Call for papers: UniReps workshop on unifying representations in neural models, NeurIPS 2025. 2025. URL: https://unireps.org/2025/call-for-papers.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character level convolutional networks for text classification. *arXiv:1509.01626*, 2015. URL: https://arxiv.org/abs/1509.01626.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. URL: https://www.cs.toronto.edu/~kriz/cifar.html.