Similarity of Neural Network Representations Revisited

Centered kernel alignment for comparing learned representations

Paper: Kornblith, Norouzi, Lee, and Hinton
Independent research reading

ICML 2019 | arXiv:1905.00414v4
20 September 2026

The question is which structure two networks share

- Feed the same n examples, in the same order, to two networks.
- Compare $X \in \mathbb{R}^{n \times p_1}$ and $Y \in \mathbb{R}^{n \times p_2}$ after column centering.
- Can we recover corresponding layers across seeds, widths, architectures, and training datasets?
- Key idea: compare relations between examples, retaining relative variance across directions.

Paper, Sec. 1 and Problem Statement, PDF pp. 1–2.

Invariance determines what similarity can see

Transformation	Linear CKA	Motivation
Neuron permutation	Invariant	Coordinate symmetry
Orthogonal rotation	Invariant	Preserve inner products
Global scaling	Invariant	Remove overall amplitude
Per-feature scaling	Sensitive	Retain relative variance
Invertible linear map	Sensitive	Avoid excessive invariance

Preserving a network's output by reparameterization does not preserve gradient-descent dynamics.

Paper, Sec. 2, PDF pp. 2–3; Table 1, p. 4.

Full sample-space rank makes strong invariance collapse

Theorem 1

For $X, Y \in \mathbb{R}^{n \times p}$ with rank n, invariance to any invertible right multiplication implies

$$s(X, Z) = s(Y, Z) \quad \text{for every } Z.$$

- Complete rows to invertible X', Y'; set $A = (X')^{-1} Y'$. Then $XA = Y$.
- Width $p \geq n$ alone is insufficient: the rank assumption matters.
- Centered data live in $\mathbf{1}^{\perp}$: apply the argument to the effective $n - 1$ dimensional sample space.

Paper, Sec. 2.1, p. 2; Appendix A, p. 12. Centering clarification: reading analysis.

Compare sample geometry instead of matching neurons

$$X, Y \longrightarrow K = XX^T,\ L = YY^T \longrightarrow \text{normalized alignment}$$

$$\langle XX^T, YY^T \rangle_F = \text{tr}(XX^T YY^T) = \|Y^T X\|_F^2.$$

- K_{ij} describes the relation between examples i and j in one network.
- Both Gram matrices are $n \times n$, even when feature widths differ.
- One scalar per layer pair; repeat across layers to obtain a heatmap.

Paper, Sec. 3, Eq. (1), PDF p. 3.

CKA normalizes a centered kernel dependence statistic

$$H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T, \qquad \mathrm{HSIC}(K, L) = \frac{\mathrm{tr}(KHLH)}{(n-1)^2}$$

$$\mathrm{CKA}(K, L) = \frac{\mathrm{HSIC}(K, L)}{\sqrt{\mathrm{HSIC}(K, K)\,\mathrm{HSIC}(L, L)}}.$$

- Equivalently: cosine similarity of HKH and HLH.
- Defined for nonzero centered Gram norms; $[0, 1]$ for PSD kernels.
- This is not mutual information; low linear CKA does not establish independence.

Paper, Sec. 3, Eqs. (3)–(4), PDF pp. 3–4; domain clarification: reading analysis.

Linear CKA preserves scale ratios without an SVD

For column-centered activations

$$\mathrm{CKA}_{\mathrm{linear}}(X, Y) = \frac{\|Y^T X\|_F^2}{\|X^T X\|_F \|Y^T Y\|_F}.$$

- Feature-space and sample-Gram computations give the same value.
- RBF alternative: $k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/(2\sigma^2))$.
- Paper sets σ to a fraction of median sample distance; this choice preserves global-scale invariance.
- Fixed absolute RBF bandwidth does not have that guarantee.

Paper, Sec. 3; Table 1, p. 4; Eq. (14), p. 5.

CCA and regression remove different amounts of scale

$$R^2_{\mathrm{CCA}} = \frac{\|Q_Y^T Q_X\|_F^2}{p_1}, \qquad R^2_{\mathrm{LR}} = \frac{\|Q_Y^T X\|_F^2}{\|X\|_F^2}.$$

- Q_X, Q_Y: orthonormal bases of the feature column spaces.
- CCA compares two whitened spaces; regression whitens the predictor side.
- SVCCA truncates principal components before CCA (threshold 0.99 in the main comparison).
- PWCCA weights canonical correlations by feature projections; it is generally asymmetric and not orthogonally invariant.

Full-column-rank notation above; use effective bases for rank-deficient inputs.

Paper, Sec. 4, pp. 4–5; Appendix C, pp. 12–14.

CKA emphasizes agreement in high-variance directions

$$R^2_{\mathrm{CCA}} = \frac{1}{p_1} \sum_{i,j} \langle u^i_X, u^j_Y \rangle^2$$

$$\mathrm{CKA}_{\mathrm{linear}} = \frac{\sum_{i,j} \lambda^i_X \lambda^j_Y \langle u^i_X, u^j_Y \rangle^2}{\sqrt{\sum_i (\lambda^i_X)^2} \sqrt{\sum_j (\lambda^j_Y)^2}}.$$

- u^i: unit eigenvector of a sample Gram matrix; λ^i: its eigenvalue.
- CCA weights subspace overlap equally; CKA weights both sides by their spectra.
- Benefit and limitation: low-variance differences contribute less.

Paper, Sec. 5, Eqs. (13)–(14), PDF p. 5.

Same subspace can mean very different sample geometry

Teaching example, not a paper experiment

Let u, v be orthonormal, zero-mean sample vectors:

$$X = [10u,\ v], \qquad Y = [u,\ 10v].$$

$$R^2_{\mathrm{CCA}} = 1, \qquad \mathrm{CKA}_{\mathrm{linear}} = \frac{100 + 100}{10000 + 1} \approx 0.0200.$$

The spaces coincide, but the dominant direction switches. CKA deliberately regards this as a large change.

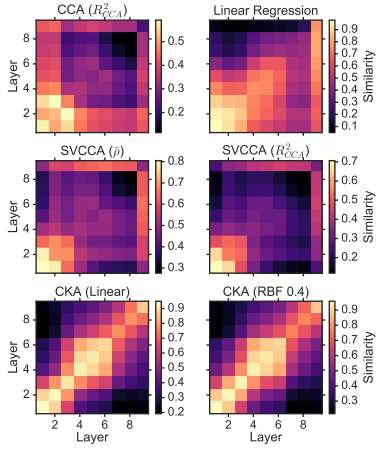
Teaching derivation from the paper's Eq. (14), p. 5.

The sanity check tests recovery of corresponding layers

- Train 10 architecturally identical Tiny-10 CNNs from different initializations.
- For each layer, find the most similar layer in each other network; exclude logits.
- Average layer-identification accuracy. Activations use the CIFAR-10 training set.
- Symmetrize PWCCA and regression as $S + S^T$.
- This operational criterion is not a universal ground truth for representation similarity.

Paper, Sec. 6.1, Fig. 2 and Table 2, PDF p. 6.

Linear CKA recovers CNN layers with 99.3% accuracy

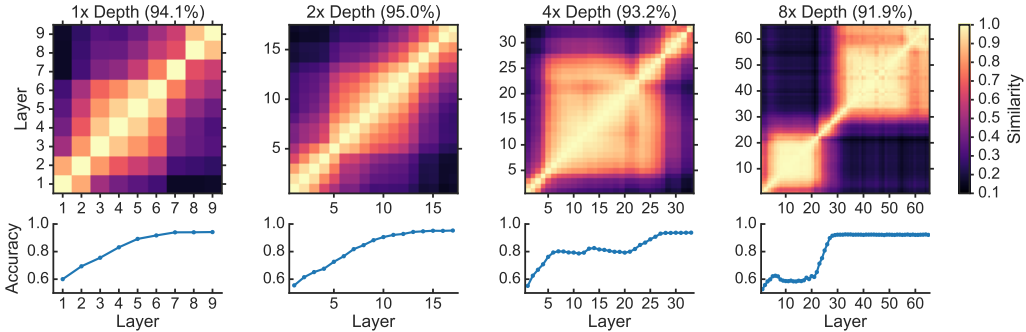


Index	Accuracy (%)
CCA, mean squared	10.6
SVCCA, mean squared	15.1
PWCCA	11.1
Linear regression	45.4
Linear HSIC	22.2
Linear CKA	99.3
RBF CKA, 0.4	99.1

Layer matching, not classification.
Linear and RBF 0.4 CKA are not significantly different under the paper's test.

Paper, Fig. 2 and selected rows of Table 2, p. 6; jackknife z-test, $p < 0.05$.

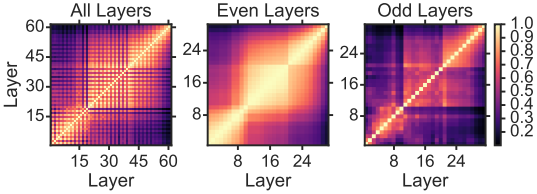
Excessive depth can stop refining representations



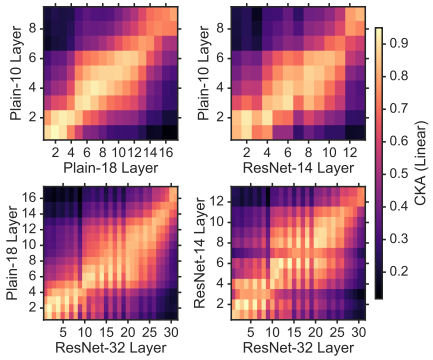
Upper: within-network CKA. Lower: layer-wise logistic probes.
The 8x-depth network plateaus before halfway through.

Paper, Sec. 6.2, Fig. 3, PDF p. 7. Titles show individual-model accuracies.

Residual structure and depth shape layer correspondence



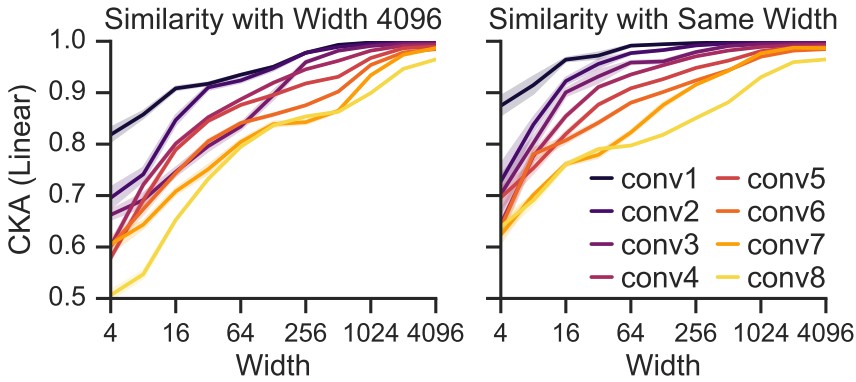
ResNet-62: post-residual activations differ from block interiors.



Cross-architecture bands suggest intermediate layers in deeper models.

Paper, Sec. 6.2, Figs. 4–5, PDF p. 7; CIFAR-10 test inputs.

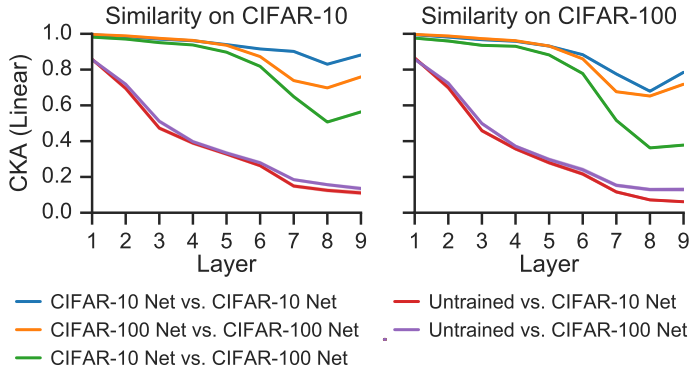
Wider networks agree more; early layers saturate first



Left: comparison with width 4096. Right: same-width models.
Colors denote conv1–conv8; bands are jackknife standard errors.

Paper, Sec. 6.2, Fig. 6, PDF p. 7; 10 models, CIFAR-10 test inputs.

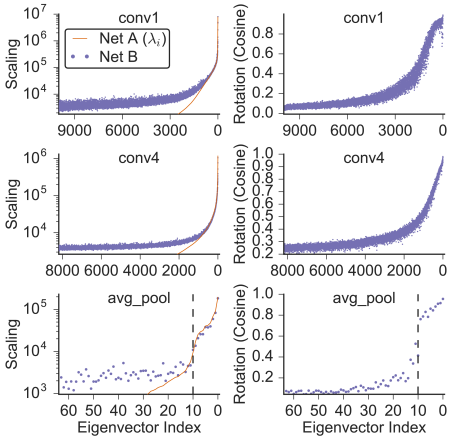
Training on different datasets still aligns early layers



Each comparison feeds identical inputs to both models. Left/right panels use CIFAR-10/CIFAR-100 test inputs.

Paper, Sec. 6.3, Fig. 7, PDF p. 8; trained and untrained controls.

Shared directions are concentrated in the large spectrum



Paper, Sec. 6.4, Fig. 8, PDF p. 8.

- Left: compare scaling by XX^T and YY^T on u_X^i.
- Right: cosine of the rotation under YY^T.
- Large eigenvalues are toward the right of each panel.
- The 64-unit pooling layer shares roughly 10 dimensions, near the class count.

An observation, not a universal rank theorem.

Transformer results depend on where activations are taken

Index	Layer Norm	Scale	Attn/FFN	Residual
CCA, mean squared	87.8	87.8	95.3	95.2
PWCCA	88.5	88.9	96.1	87.0
Linear CKA	78.6	95.6	86.0	73.6
RBF CKA, 0.4	92.3	96.5	89.1	98.1

- Layer-identification accuracy (%); all tested methods are non-trivial.
- CCA ignores channel-wise scaling; CKA can reveal its consistent structure.
- Linear CKA is not uniformly best, notably after residual addition.

Paper, Appendix F.1, selected rows of Table F.1, p. 18; no significance markings reproduced.

Control experiments narrow the interpretation

- Alternative SVCCA thresholds still miss the tested CNN layer structure.
- Untrained networks also exhibit architecture-induced similarity; high CKA alone does not establish shared learned knowledge.
- With and without batch normalization, representations remain related; hyperparameters were tuned separately.
- Within-class CKA resembles all-example CKA, so between-class separation is not the only source of structure.

Paper, Appendix F.2–F.4, Figs. F.3–F.7, PDF p. 19.

CKA is useful evidence, not functional equivalence

- Interpretation depends on the input sample, kernel, and exact activation location.
- High-variance weighting can underemphasize low-variance task-relevant differences.
- Avoiding the invariance theorem does not guarantee freedom from finite-sample or high-dimensional estimation effects.
- Heatmap similarity does not establish causal interchangeability, safe pruning, or equal out-of-distribution behavior.

Paper, Sec. 7, p. 8; evidence boundaries and application implications: reading analysis.

Preserve sample geometry and report the conditions

1. Choose invariances to match the question: CKA retains relative spectral scale.
2. Read the strong CNN correspondence result together with the Transformer exceptions.
3. Pair similarity with probes, spectral inspection, and interventions when making functional claims.

Discussion: Which differences matter for your task even when their variance is small?

Synthesis of Secs. 2–7 and Appendix F; final question is a discussion prompt.

Reference and reading provenance

Similarity of Neural Network Representations Revisited
Simon Kornblith, Mohammad Norouzi, Honglak Lee, Geoffrey Hinton.
ICML 2019; arXiv:1905.00414v4 (19 July 2019).

`https://arxiv.org/abs/1905.00414v4`

Read from the matching author TeX source and 20-page PDF.
Main text and Appendices A–F covered.
Figures reproduced from author assets and checked against the PDF.
Prepared as an independent reading presentation on 20 September 2026.

Page references in this deck use 1-based PDF file pages.

Backup: regularization connects the similarity family

$$\widetilde{Q}_X = U_X \Sigma_X (\Sigma_X^2 + \kappa_X I)^{-1/2}$$

$$\|\widetilde{Q}_Y^T \widetilde{Q}_X\|_F^2 = \sum_{i,j} \frac{\lambda_X^i \lambda_Y^j}{(\lambda_X^i + \kappa_X)(\lambda_Y^j + \kappa_Y)} \langle u_X^i, u_Y^j \rangle^2.$$

- Zero regularization: recover CCA with its corresponding normalization.
- One-sided large regularization: a suitable normalization recovers regression.
- Both sides large: separable normalization converges to linear CKA.

Do not mix the normalization choices when taking these limits.

Paper, Appendix D.1, PDF pp. 14–16.

Backup: a dimensional check on the regression identity

$$\mathrm{CKA}(XX^T, Q_Y Q_Y^T) = \frac{\|Q_Y^T X\|_F^2}{\|X^T X\|_F \sqrt{p_2}}$$

$$R^2_{\mathrm{LR}} = \mathrm{CKA}(XX^T, Q_Y Q_Y^T) \frac{\sqrt{p_2}\|X^T X\|_F}{\|X\|_F^2}.$$

- The paper's Eq. (12) prints $\sqrt{p_1}$ in both PDF and TeX.
- With $Q_Y^T Q_Y = I_{p_2}$, direct substitution gives $\sqrt{p_2}$: an apparent typo.
- No difference for equal widths; use rank(Y) for an effective rank basis.

Reading derivation checking Paper, Eq. (12), PDF p. 5; not an author-issued correction.

Backup: protocols and accuracy numbers are not interchangeable

- Fig. 3 individual Plain models: 94.1, 95.0, 93.2, 91.9% classification accuracy.
- Table E.2 means: 93.9, 94.8, 93.7, 91.3% for Plain-10/18/34/66.
- Transformer: WMT18 English–German; 10 encoders (45 pairs); 12 alternating attention/FFN sublayers.
- Evaluation: 3,000 newstest2013 sentences, 75,804 tokens, excluding end-of-sentence tokens.
- Main CNN matching and Appendix F.5 use training inputs; architecture/width plots use test inputs.

Paper, Fig. 3, p. 7; Table E.2, p. 17; Appendix F.1, p. 18; F.5, p. 20.