Bridging Critical Gaps in Convergent Learning: How Representational Alignment Evolves Across Layers, Training, and Distribution Shifts

Chaitanya Kapoor, Sudhanshu Srivastava, Meenakshi Khosla

Department of Cognitive Science, UC San Diego, {chkapoor, sus021, mkhosla}@ucsd.edu

Abstract

Understanding convergent learning—the degree to which independently trained neural systems—whether multiple artificial networks or brains and models—arrive at similar internal representations—is crucial for both neuroscience and AI. Yet, the literature remains narrow in scope—typically examining just a handful of models with one data distribution, relying on one alignment metric, and evaluating networks at a single post-training checkpoint. We present a large-scale audit of convergent learning, spanning dozens of vision models and thousands of layer-pair comparisons, to close these long-standing gaps. First, we pit three alignment families against one another—linear regression (affine-invariant), orthogonal Procrustes (rotation-/reflection-invariant), and permutation/soft-matching (unit-order-invariant). We find that orthogonal transformations align representations nearly as effectively as more flexible linear ones, and although permutation scores are lower, they significantly exceed chance, indicating a privileged representational basis. Tracking convergence throughout training further shows that nearly all eventual alignment crystallizes within the first epoch—well before accuracy plateaus—suggesting it is largely driven by shared input statistics and architectural biases, not by convergence towards the final task solution. Finally, when models are challenged with a battery of out-of-distribution images, early layers remain tightly aligned, whereas deeper layers diverge in proportion to the distribution shift. These findings fill critical gaps in our understanding of representational convergence, with implications for neuroscience and AI.

Keywords: Representational Convergence

Introduction

Deep Neural Networks (DNNs) are becoming indispensable tools in neuroscience, for both predicting neural responses (Yamins et al., 2014), (Yamins & DiCarlo, 2016), (Khaligh-Razavi & Kriegeskorte, 2014) or as models for reverse-engineering algorithms of neural computation (Schrimpf et al., 2018), (Schrimpf et al., 2020), (Cichy et al., 2016). This congruence invokes the necessity to gain a deep understanding of how DNNs learn to represent information. A core question in this domain is whether—and under what conditions—independently trained networks develop similar internal representations, and along which dimensions this convergence unfolds. Answering question along these lines of inquiry can help illuminate how architectural design, training objectives and input statistics give rise to emergent features, and whether any aspects of these representations are universal across architectural choices.

Over the past decade, studies have shown that independently trained models often learn highly aligned feature sets—early convolutional layers reliably produce Gabor-like filters (Yosinski et al., 2014), and deeper layers exhibit substantial correspondence when measured via methods such as CCA and its variants (Morcos et al., 2018; Raghu et al., 2017) CKA (Kornblith et al., 2019), RSA (Mehrer et al., 2020) and model stitching (Bansal et al., 2021). More recently, large-scale models appear to converge even across tasks and data modalities (Moschella et al., 2023), suggesting the existence of a modality-agnostic, *"Platonic representation"* that transcends specific inputs (Huh et al., 2024). Yet, these findings leave open critical gaps: we lack fine-grained metrics to characterize the minimal transformations needed for alignment; we know little about how and when representations align *during* training; and we do not understand how robust this convergence is under distributional shifts. In this work, we address these questions by evaluating representational alignment along three key axes—across layers, over the course of training, and under out-of-distribution (OOD) inputs—using a suite of metrics with varying degrees of invariances to reveal the geometry and dynamics of convergence in DNNs.

Results

Consider a representational pair $\boldsymbol{X}_i \in \mathbb{R}^{M \times N_x}$ and $\boldsymbol{X}_j \in \mathbb{R}^{M \times N_y}$ as responses to M unique stimuli, where N_x and N_y denote the number of units in each representation. We compare these under three mappings—permutation (and its extension when $N_x \neq N_y$, Soft-Matching), Procrustes and Linear regression, by finding the optimal mapping $\boldsymbol{M}$ under each transformation. We then report the alignment as a pairwise correlation $\texttt{corr}(\boldsymbol{X}_j, \boldsymbol{M}\boldsymbol{X}_i)$. These metrics, ordered from strict to flexible, isolate alignment in tuning functions, geometric shape and information content respectively.

Convergence with Network Depth: Across independent seeds of the same architecture, alignment is highest in the early layers—known to extract universal, low-frequency features (*e.g.,* edges, corners) (Rahaman et al., 2019; Bau et al., 2017; Zeiler & Fergus, 2014)—and gradually decreases with network depth, consistent across all three metrics (Fig. 1-A). Allowing more flexible mappings (Permutation $\rightarrow$ Procrustes $\rightarrow$ Linear) yields marginal gains beyond Procrustes, indicating that simple rotations capture most of the shared structure and additional degrees of freedom (*e.g.,* scaling, shearing) contribute little. Since Procrustes is symmetric, these results

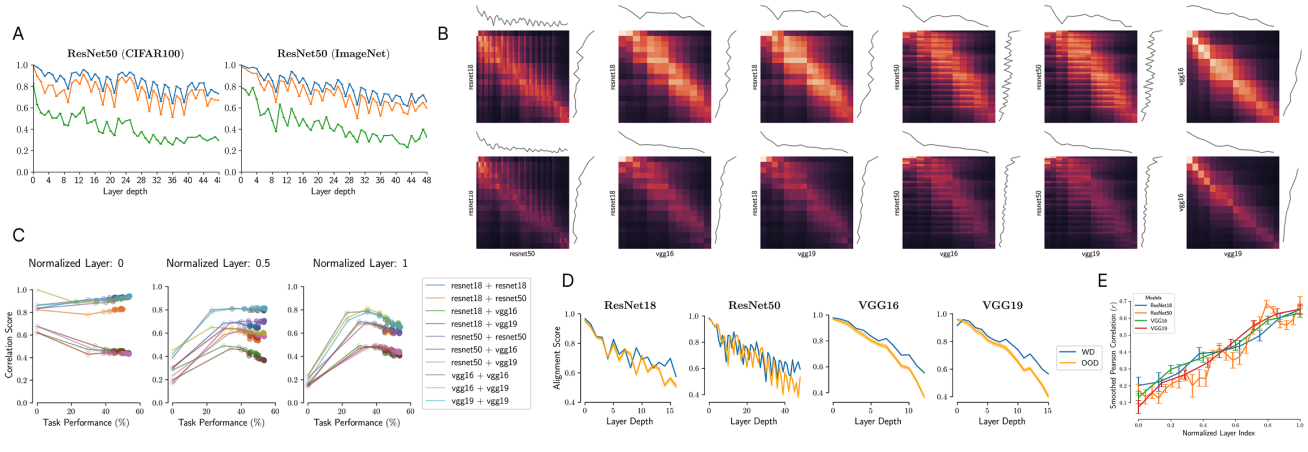

Figure 1: **Representational Alignment Across Network Depth, Training and Distribution Shifts. (A)** Alignment score vs. layer depth, showing stricter metrics yield lower scores (Linear > Procrustes > Permutation). **(B)** Layer-layer alignment using Procrustes (**Top**) and Soft-Matching (**Bottom**) with maximums over rows (top) / columns (right) denoted in gray lines. **(C)** Procrustes alignment over the first 10 ImageNet training epochs (lighter = earlier), with task performance. **(D)** Layer-wise Procrustes alignment for within-distribution (WD) and 17 OOD datasets. Error bars denote standard error across $(n = 17)$ OOD sets. **(E)** Correlation between Procrustes alignment and task performance over normalized network depth.

highlight that convergent learning reflects a deep geometric similarity in feature organization across networks, not merely the ability of one network to predict another.

Hierarchical Correspondence holds Across Metrics: Previous studies have shown that, for architecturally identical networks trained from different initializations, the most similar layer in one network to a given layer in another is the corresponding architectural layer (Kornblith et al., 2019). However, this finding has primarily been supported using affine-invariant metrics (*e.g.,* CCA, SVCCA). We extend this result by showing that stricter metrics—such as Procrustes and soft-matching scores—also reveal the same hierarchical correspondence (Figs. 1-B), even when comparing networks with different architectures. This suggests that the hierarchical alignment of representations is a fundamental property of neural networks, robust to architectural differences. Moreover, our results show that both representational shape (Procrustes) and neuron-level tuning (Soft-Matching) follow similar alignment patterns, reinforcing the consistency of this hierarchical organization across different levels of representational analysis.

Evolution of Convergence Over Training: We measure Procrustes alignment between independently trained ImageNet networks over the first ten epochs and found that the bulk of convergence occurs within the first epoch—well before appreciable task accuracy (Fig. 1-C). This rapid early convergence suggests that shared input statistics, architectural inductive biases, and initial training dynamics are primary drivers of alignment, rather than the final task-specific solution. Such findings stand in contrast to hypotheses such as the contravariance principle (Cao & Yamins, 2021) and *task generality hypothesis* (Huh et al., 2024), which attribute convergence to constraints imposed by the final, high-performance solution. These results are consistent with studies on the early training phase that identify a rapid representational reorganization before meaningful task learning begins (Frankle et al., 2020).

Convergence Across Distribution Shifts: We investigate the extent of within-distribution (WD) representational alignment of ImageNet-trained DCNNs under 17 OOD variants of ImageNet (Geirhos et al., 2018) sharing ImageNet's 16 coarse labels by computing layerwise Procrustes scores. Early convolutional layers maintained alignment levels nearly identical to WD stimulus, whereas later layers exhibit amplified divergence under OOD conditions (Fig. 1-D). We attribute this to early layers encoding universal features (*e.g.,* edges, corners) that generalize across distributions, while deeper layers represent task-specific abstractions that are more sensitive to distributional shifts. Further, we find that representational alignment in deeper (but not earlier or middle) layers correlates with OOD classification accuracy (Fig. 1-E). These findings suggest that OOD stimuli can serve as an effective probe to differentiate model architectures in model–brain comparisons, and that selective fine-tuning of later layers can serve as a promising strategy for improving OOD generalization.

Discussion

This study fills critical gaps in our understanding of convergent learning, offering a comprehensive analysis of how representational alignment between independently trained networks varies across network depth, training, and distribution shifts. We systematically explored how different alignment metrics—with varying levels of transformation invariance—capture representational similarities, providing a more nuanced view of convergent learning than previous work.

References

Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, *34*, 225–236.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6541–6549).

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*(1), 27755.

Frankle, J., Schwab, D. J., & Morcos, A. S. (2020). The early phase of neural network training. *arXiv preprint arXiv:2002.10365*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature communications*, *11*(1), 5725.

Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, *31*.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., & Rodolà, E. (2023). Relative representations enable zero-shot latent space communication (2023). *arXiv preprint arXiv:2209.15430*.

Raghu, M., Gilmer, J., Yosinski, J., & Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, *30*.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., ... Courville, A. (2019). On the spectral bias of neural networks. In *International conference on machine learning* (pp. 5301–5310).

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part i 13* (pp. 818–833).