Shared Representational Geometry Across Neural Networks

Qihong Lu[1], Po-Hsuan (Cameron) Chen[1,2], Jonathan W. Pillow[1], Peter J. Ramadge[1], Kenneth A. Norman[1], Uri Hasson[1]

1. Princeton University; 2. Google Brain

Questions

When many instantiations of the same neural network architecture are trained on the same dataset, they tend to represent highly similar mathematical functions with very different weight configurations. In what sense are these networks similar? What is the connection across these neural networks?

Method: SRM and RSM

Shared Response Model (SRM)

- SRM has the following objective:

$$\min_{\mathbf{W_i},\mathbf{S}} \sum ||\mathbf{X_i} - \mathbf{W_i}\mathbf{S}||_F^2 \text{ s.t. } \mathbf{W_i}^T\mathbf{W_i} = \mathbf{I_k} \quad (1)$$

- Geometrically speaking, we used SRM to find orthogonal transformations (k = #units) to align activity patterns across networks to a shared feature space [1-2] (Fig. 1).

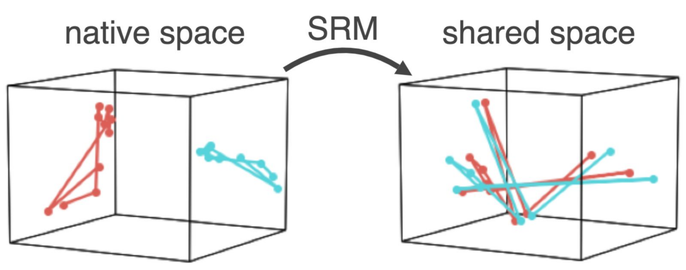

Figure 1: The hidden representations from two networks (red and cyan) before vs. after alignment.
*Results from the ConvNets-CIFAR10 experiment.

Representational Similarity Matrix (RSM)

- Within-network RSM, or **wRSM**: the correlation between the evoked neural responses for two stimuli **within** a network.
- Inter-network RSM, or **iRSM**: the correlation between the averaged neural response to a stimulus for N-1 networks vs. the response to a stimulus for a **held-out** network.

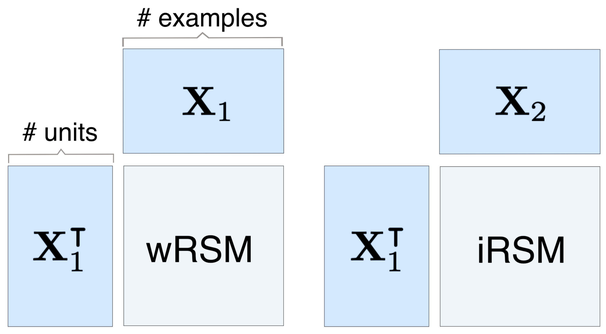

Figure 2: Demo: within-network RSM (wRSM) and inter-network RSM (iRSM) in the two-network case. $\mathbf{X}_i$ is the neural activity matrix for the i-th network.

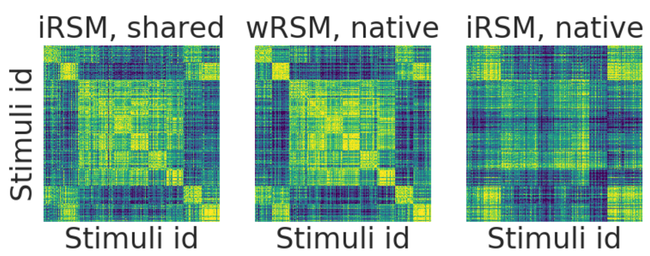

Figure 3: In the shared space (after alignment), inter-network RSM (iRSM) is similar to within-network RSM (wRSM), whereas in native spaces, iRSM does not reflect meaningful structure. *Results from the ConvNets-CIFAR10 experiment.

Experiment, align ConvNets/ResNets

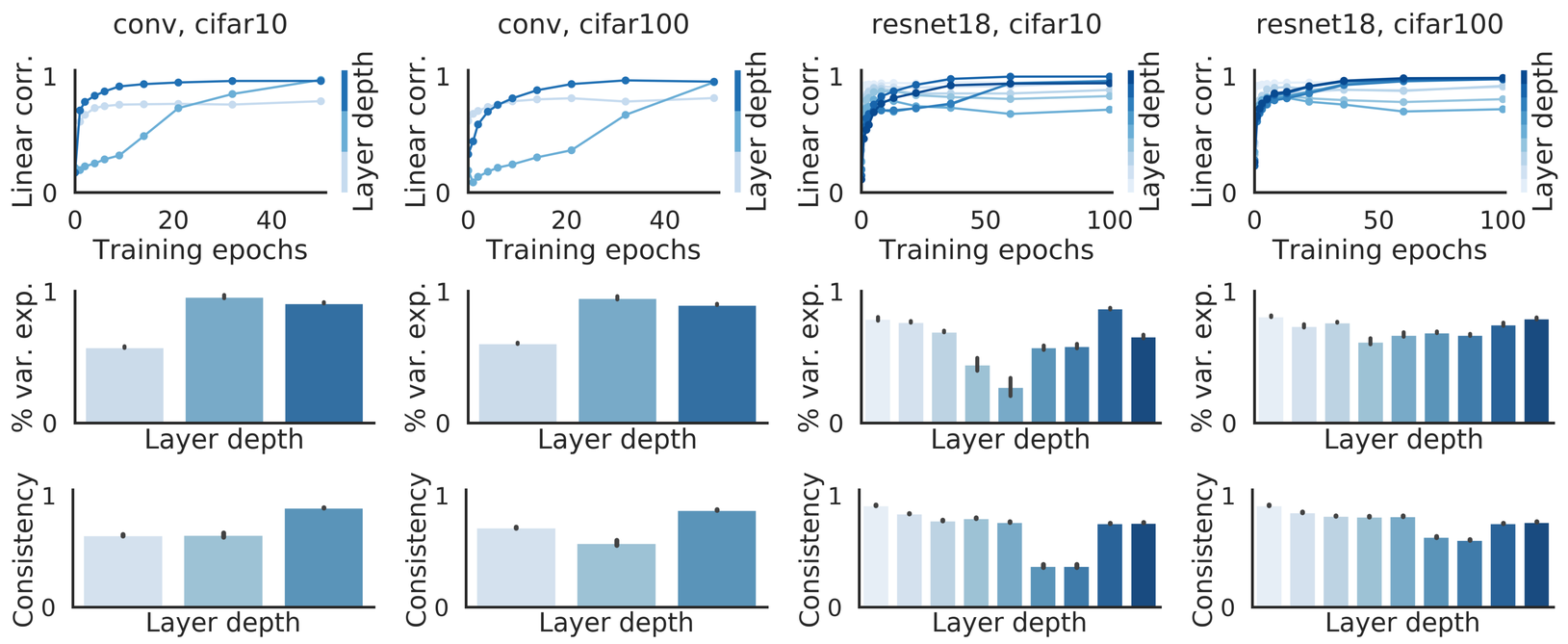

Figure 4: After training, the shared iRSM became highly similar to the converged wRSM, indicating good inter-network alignment. Top) Correlation between shared iRSM vs. the final wRSM. Mid) Variance explained by SRM, for all layers. Bot) Correlation between wRSMs across networks, for all layers. *All errorbars indicate 95% bootstrap confidence interval.

Results:

- Different trained ConvNets were well aligned by orthogonal transformations, suggesting they learned qualitatively the same hidden representation (i.e. same geometry).
- Orthogonal transformations also explained a large amount of variance for ResNets, though the alignment is not as good as ConvNets, suggesting high capacity models might learn qualitatively different representation.
- Our results are consistent with prior empirical works [3-5] and theoretical works [6].

Main points

- Approximately speaking, different neural networks learned different orthogonal transformation of the same representation.
- This consistency of representational geometry [7] across networks came from their shared experience.

Simulation: SRM can undo orthogonal transformations

- We trained a neural net, then transformed its activity patterns by some random orthogonal matrices.
- We used SRM to align different transformed activity patterns. If the alignment is perfect, shared iRSM should be the same as wRSM.

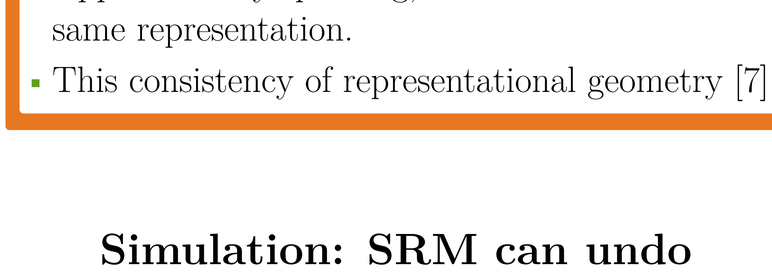

Figure 5: The average correlation between averaged wRSM vs. iRSM in the shared space (blue) and the native space (brown).

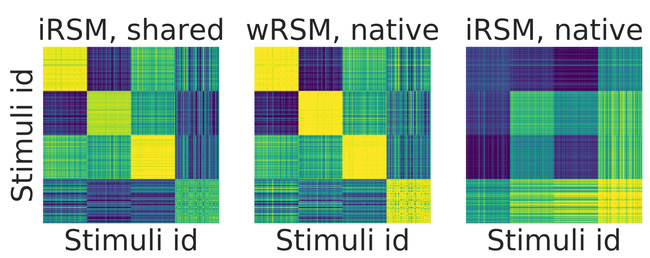

Figure 6: Shared iRSM is highly similar to the wRSM.

Neural/Cognitive models of inter-subjects correlation (ISC)

Functional alignment for NNs enabled modeling of group-level neuroimaging results, such as brain synchronization during naturalistic processing [8].

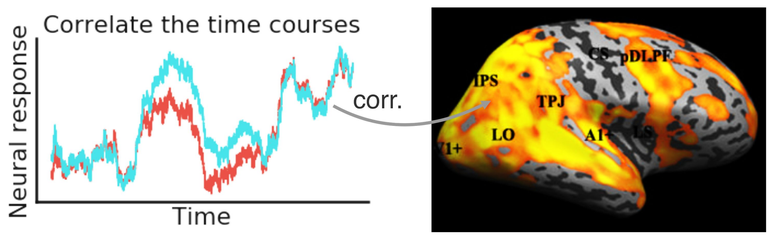

Figure 7: ISC map during movie watching (adapted from [8]).

References & Links

[1] Chen, et al. NeurIPS 2015
[2] Haxby, et al. Neuron 2011
[3] Li, et al. ICLR 2015
[4] Raghu, et al. NeurIPS 2017
[5] Morcos, Raghu, & Bengio. NeurIPS 2018
[6] Saxe, McClelland, & Ganguli. Arxiv 2018
[7] Kriegeskorte, & Kievit. Trends Cogn Sci 2013
[8] Hasson, Malach, & Heeger. Trends Cogn Sci 2010

- Download this poster: `https://tinyurl.com/nnsrm-NeurIPS18`
- Demo: `https://qihongl.github.io/nnsrm-NeurIPS18.html`