Representational similarity across neural networks

Qihong Lu[1], Po-Hsuan (Cameron) Chen [2,3], Peter J. Ramadge[2], Kenneth A. Norman[1], Uri Hasson[1]

[1]Department of Psychology and Princeton Neuroscience Institute, Princeton University; [2]Department of Electrical Engineering, Princeton University; [3]Google Brain

Summary

Different neural networks can learn to represent highly similar mappings using different connection weights. We found hyperalignment can identify their shared representational geometry. Moreover, the shared representational geometry converges to the within-subject representational geometry. This suggests that different neural networks learn highly similar internal representations.

Method

Shared response model (SRM)[1]

- idea: align representational geometries across subjects to a shared feature space by (roughly) rigid-body transformations (Fig. 2).
- detail: given neural data $\mathbf{X_i}$, find $\mathbf{S}$, the shared time course, and $\mathbf{W_i}$, the subject-specific transformation matrices with the following objective:

$$\min \sum_i ||\mathbf{X_i} - \mathbf{W_i}\mathbf{S}||_F^2 \quad \text{s.t.} \quad \mathbf{W_i}^T\mathbf{W_i} = \mathbf{I_k}$$

Representational similarity Analysis[4]

- **within-subject**: the correlation between the evoked neural responses for two stimuli.
- **inter-subject**: the correlation between the averaged neural response to a stimulus from N-1 subjects vs. the response to a stimulus from a held-out subject.

References & Acknowledgement

[1] Chen, P.-H., et al. (2015). NIPS 2015.
[2] Hasson, U., et al. (2004). Science.
[3] He, K., et al. (2016). CVPR 2016.
[4] Kriegeskorte, N. et al. (2008). Front Syst Neurosci.
*SRM code is in BrainIAK: `http://brainiak.org/`

Acknowledgement: This work was supported by a Multi-University Research Initiative grant to KAN and UH (ONR/DoD N00014-17-1-2961). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research or the U.S. Department of Defense.

Representational similarity

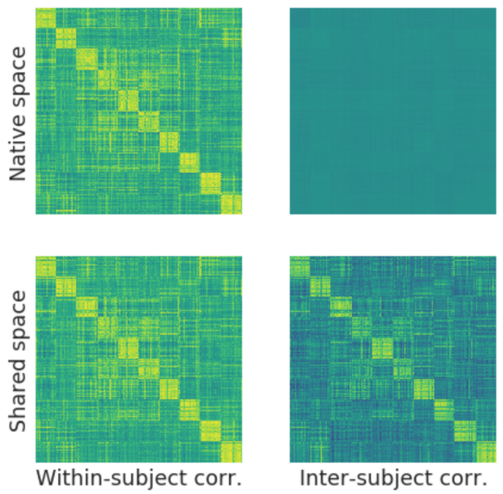

Figure 1: After training, the inter-subject RSM in the shared space is similar to the within-subject RSM.

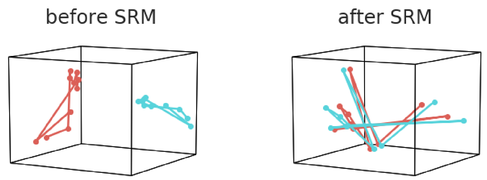

Figure 2: SRM can align representational geometries.

Simulation: Discover shared representation

- We trained a neural net, then transformed its representation by some random orthogonal matrices.
- We used SRM to reconstruct the original representation.

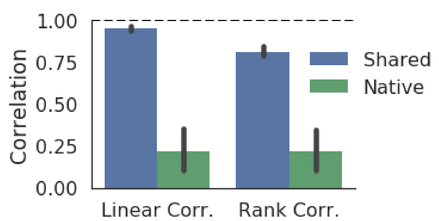

Figure 3: SRM can discover shared representational geometry, and it is invariant under orthogonal perturbations.

Developmental data

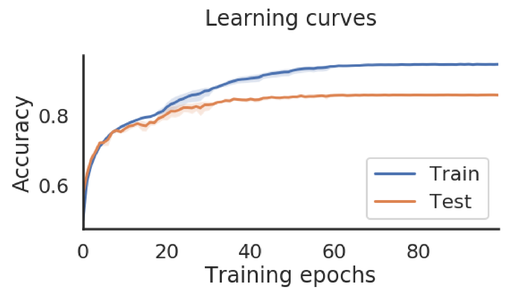

Figure 4: We trained 8 residual networks[3] on CIFAR10. Each residual network has 18 layers.

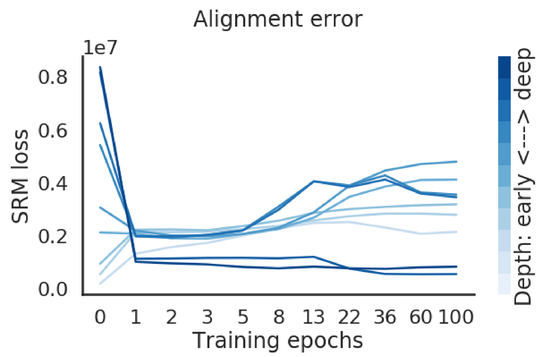

Figure 5: Over the time course of training, early layers got slightly less aligned; deeper layers got more aligned.

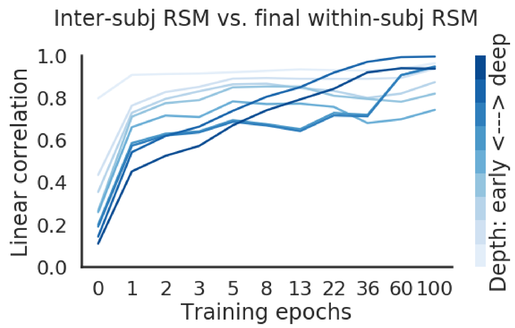

Figure 6: Training made the inter-subject RSM more similar to the final within-subject RSM. This shows different neural networks converge to the same representational geometry.

Computational modeling of ISC

- Inter-subject correlation (ISC): People's brain dynamics can be synchronized by a common stimulus[2]. This is measured by correlating voxel time courses across subjects (Fig. 7).
- We can compute ISC in the shared space across neural nets, which enabled us to model the emergence of ISC.

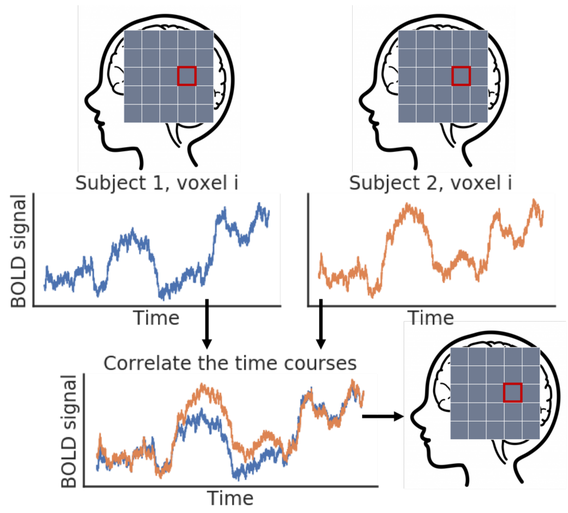

Figure 7: A schematic diagram of the ISC pipeline.

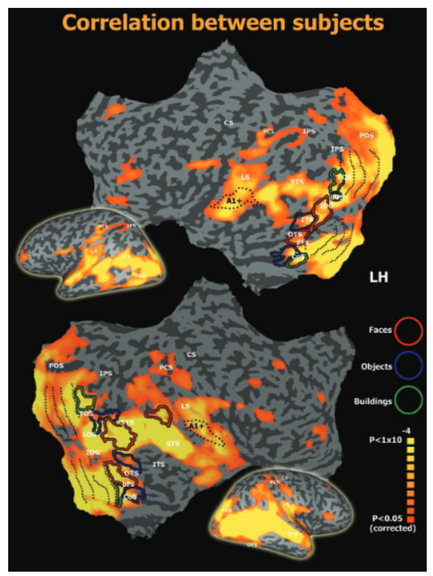

Figure 8: An ISC map while people were viewing a movie [2].