

Cognitive Neuroscience and Deep Learning

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 22 of "The NeuroAI Handbook"

Using deep networks as testable theories of the brain: representations, RSA, and the visual hierarchy.

One-sentence thesis. A deep network trained on the brain's own task becomes a testable theory of the brain, and you test it not by comparing wiring but by comparing the geometry of what the two systems represent.

PART 1: A Testable Brain Theory

From Metaphor to Model

For centuries neuroscience described the brain with metaphors: a switchboard, a logic engine, a computer. These were just words. We had no working machine that could actually *do* what a brain region does and whose insides we could then inspect.

Deep learning ended the drought. We now have models that recognize objects and understand language at human-rivaling levels, which turns a metaphor into a scientific claim:

The core hypothesis. a goal-driven deep network,

trained to solve the same problem a brain region solves, is a *candidate model* of how that region computes.

The word candidate matters, this is a testable hypothesis, not a poem. If a network trained to recognize objects lands on the same solution the brain uses, we can treat the model as a **computational microscope**: dissect its layers to form concrete, checkable guesses about the brain's algorithm. The catch is that you cannot open both systems and line up the parts. Wetware and software share no components. So what, exactly, do you compare?

The Right Thing to Compare

You compare **representations**, not hardware. A representation is the pattern of activity across a population of units (neurons or model features) evoked by a stimulus. The insight is to ignore the individual units entirely and ask about *geometry*: which stimuli produce similar patterns, and which produce different ones? Two systems can be built from utterly different parts yet organize the same stimuli the same way.

ANALOGY

To learn whether Alice and Bob think about movies alike, do not read their minds, hand them the same 50 films and watch how they shelve them. Alice sorts by genre, Bob by director. You never inspect a single neuron, yet the two arrangements reveal whether their mental geometry agrees.

same stimuli to a brain (via fMRI or electrophysiology) and to a model. For each system, measure the pairwise dissimilarity of the evoked patterns and pack them into a **representational dissimilarity matrix (RDM)**.

Formally, one cell of the RDM is one minus the correlation between two stimuli's patterns:

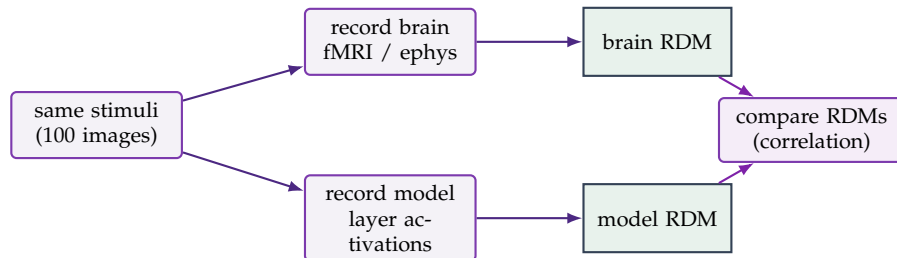
$$\text{RDM}_{ij} = 1 - \text{corr}(\mathbf{r}_i, \mathbf{r}_j)$$

The RDM is a fingerprint of the representational geometry that no longer references any particular neuron, so a brain RDM and a model RDM are directly comparable. The **RSA score** is simply how well the two fingerprints agree:

$$\text{score} = \text{corr}(\text{RDM}_{\text{brain}}, \text{RDM}_{\text{model}})$$

RSA, Step by Step

Representational Similarity Analysis (Kriegeskorte et al., 2008) formalizes the shelving test. Show the



The RSA pipeline. Because the RDM abstracts away the units, an fMRI voxel pattern and a network layer can be compared on equal footing.

A Hierarchy in the Brain

The clearest win came from vision. The **ventral visual stream**, the brain's object-recognition pathway, is a known hierarchy: **V1** responds to simple edges, **V4** to shapes and textures, and **inferotemporal (IT) cortex** to whole objects. Complexity and invariance grow at each stage.

A Hierarchy in the Network

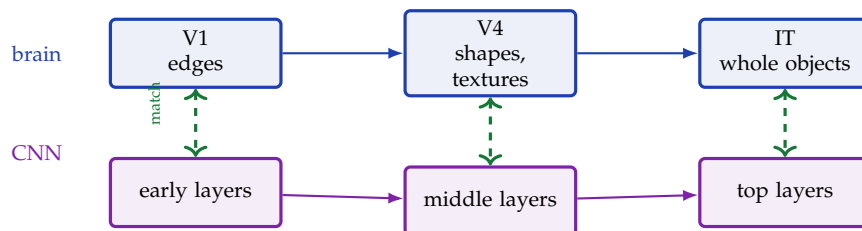
Train a **convolutional neural network** (AlexNet, VGG) on object classification and something remarkable falls out: its layers line up with the brain's stages. Early layers match V1, middle layers match V4, top layers match IT (Yamins and DiCarlo, 2014; Guclu and van Gerven, 2015). A system optimized for one goal, classify the object, independently discovered the

same cascade of hierarchical feature extraction that evolution built.

Why this is big. the model predicts not just *what* the brain represents but offers a reason *why*, because that hierarchy is an efficient solution to the recognition problem. Two very different optimizers, evolution and gradient descent, converged on it.

ANALOGY

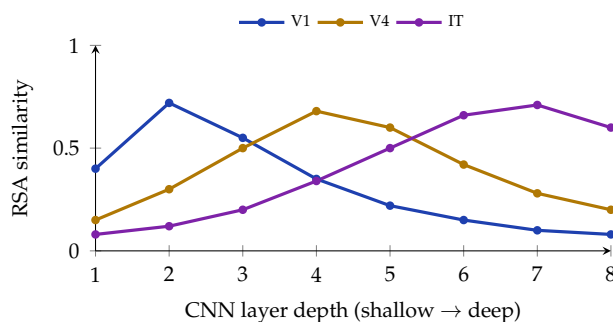
The camera-like eye evolved separately in vertebrates and octopuses because a lens over a sheet of photoreceptors is simply a good solution to seeing. When a CNN and the visual cortex land on the same layered features, it is that kind of convergence, the problem shaped the answer.



Layer-by-layer correspondence: each stage of the ventral stream peaks in similarity with a specific CNN layer, in the expected order.

Plotted against layer depth, each brain region's similarity to the network peaks at a different layer, early

regions early, late regions late, tracing the diagonal that gives the correspondence away.



Each region's similarity curve peaks at a different depth, V1 shallow, IT deep. The staggered peaks are the fingerprint of a shared hierarchy.

PART 4: What It Buys Neuroscience

From “Where” to “How”

For decades imaging answered *where*: face processing lives in the fusiform face area, and so on. A location is not a mechanism. A working model answers *how*, because you can open it, read out each layer, and turn “this circuit does X” into a concrete, testable algorithmic claim.

Playing What-If

A model that predicts brain activity becomes an experimental sandbox. **Lesion** units and see whether the failures mirror patients with brain damage. Run **optimal-stimulus synthesis**, search for the image that maximally drives a target unit, then present it to a real neuron to check the prediction. In Marr’s terms, these models bridge the **computational** level (what prob-

lem is solved) and the **algorithmic** level (how), even where the **implementation**, neurons versus silicon, differs. Benchmarks like **Brain-Score** (Schrimpf et al., 2018) turn “how brain-like is this network?” into a number you can climb.

The Limits Are the Point

Do not oversell it. CNNs also diverge from brains, they crumble under adversarial pixels, need far more examples to learn, and lack the rich recurrence of cortex. Those discrepancies are not embarrassments, they are the next hypotheses. Where model and brain part ways is exactly where the science gets interesting, and it is fueling a genuinely bidirectional loop: neuroscience inspires better AI, and AI gives neuroscience testable theories.

Quick Self-Test

1. Why can’t you test the core hypothesis by comparing neurons to model units directly, and what do you compare instead?
2. Walk through the four steps of RSA, from a shared stimulus set to a single similarity score.
3. What does an RDM capture, and why does abstracting away the individual units make brain and model comparable?
4. State the ventral-stream to CNN correspondence, and explain why “convergence” is the right word for it.
5. Name one way CNNs differ from the visual cortex, and argue why that difference is scientifically useful.

If You Remember One Thing

The move that turned deep networks into brain science is comparing *representational geometry*, not wiring. Build an RDM for a brain region and one for a model layer, correlate them, and a CNN trained only to name objects turns out to organize vision the way V1, V4, and IT do, evidence that the problem, not the substrate, shaped the solution.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552