Wharton
UNIVERSITY of PENNSYLVANIA
DEPARTMENT OF STATISTICS
AND DATA SCIENCE

Representations in Deep Neural Networks

What they are, why we care, and how we can use them

Joseph H. Rudoler

February 20, 2025

Motivation

We talk a lot in ML/AI about "representations" but the idea is either fuzzy / circular ("something useful", "high-dim representation") or overly technical/concrete ("activations of hidden layers")

This talk will hopefully cover:

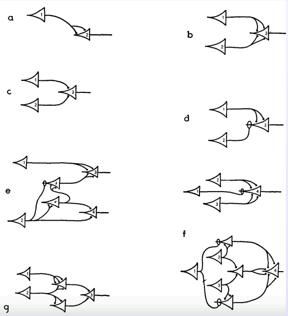
- Intuition for how representations in ANNs were developed
- Theoretical frameworks
- Survey modern use cases

Representations in Neuroscience

Individual neurons fire in an "all or nothing" manner.

- This 0-or-1 activity can be treated as a boolean primitive
- A connected network of neurons can express many complex logical functions.[a]

[a] McCulloch and Pitts, 1943.



Representations in Neuroscience

Neural networks are capable of learning

- Discrete boolean logic is too rigid
- Hebbian learning[a] framework explained how co-activation could strengthen connections
  - *Neurons that fire together, wire together*
- Extension to probabilistic setting produced the "perceptron"[b], with a convergence theorem in the case of linearly separable classes

[a] D.o Hebb, 1949.
[b] Rosenblatt, 1958.

64 *Growth of the Assembly*

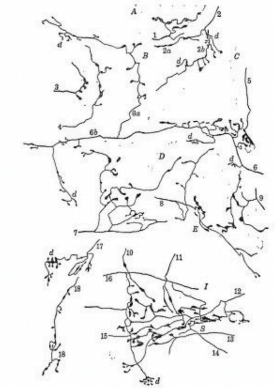


FIGURE 6. Relationships between synaptic knobs and the cell body. From Lorente de Nó, 1938*a*. Courtesy of Charles C. Thomas and of the author.

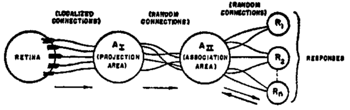


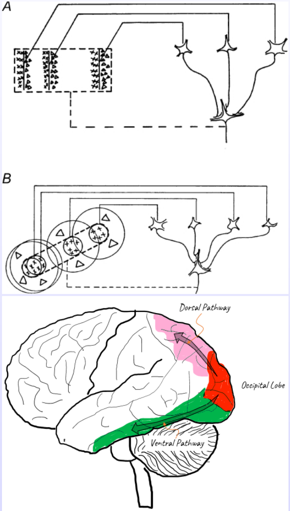
FIG. 1. Organization of a perceptron.

Hierarchical representations in neuroscience

Biological neurons in the visual cortex exhibit hierarchy[a]

- Experiments on cats - observed neural firing when shown simple images / shapes.
- (B) circular local receptive fields ⇒ "simple" cells that can identify lines/boundaries
- (A) Simple cells ⇒ complex cells
- Information propagates through visual cortex[b]

[a] Hubel and Wiesel, 1962.
[b] Goodale and Milner, 1992.



Hierarchical representations in computer vision

Convolutional Neural Networks (CNNs) were inspired by biological NNs

- First CNN developed by Fukushima, manually-designed kernels to recreate the same feature extraction[a]
- Later developed more flexible "neocognitron" using Hebbian unsupervised learning[b]

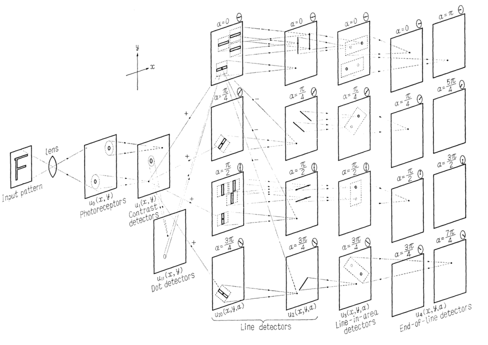


Fig. 5. Sketch of the network structure. Only typical interconnections are shown. The elements in a layer (enclosed within a solid line) are connected to the element in the next layer with positive interconnecting coefficients; those enclosed within a dotted line are connected with negative interconnecting coefficients.

---

[a] Fukushima, 1969.
[b] Fukushima, 1980.

Hierarchical representations in Computer Vision

Backpropagation was the key to learning "useful" representations

Learning representations by back-propagating errors (Rumelhart et al., 1986)

"In perceptrons, there are "feature-analysers" between the input and output that are not true hidden units because their input connections are fixed by hand, so their states are completely determined by the input vector: they do not learn representations. The learning procedure must decide under what circumstances the hidden units should be active **in order to help achieve the desired input-output behaviour**. This amounts to deciding what these units should represent."

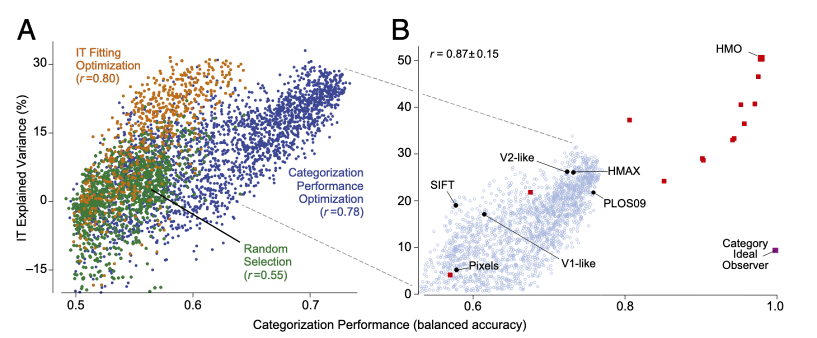
- Yann Lecun applied supervised learning to train CNN kernels via backpropagation (classic MNIST results)[1]

[1] LeCun et al., 1989.

Correspondence between visual system and CNNs

Does the brain extract the same features as CNNs?

- Yamins et al., 2014 was one of the first studies to show a correspondence between representations in brains and ANNs
- This area of research comes with *lots* of caveats



Parallel Distributed Processing (PDP)

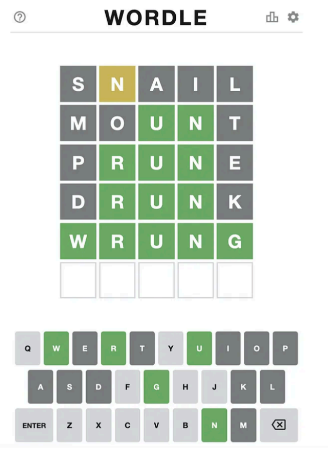
So... why are NN representations good?

- McClelland et al., 1986 posit that there are advantages to models which compose smaller "units", processed in parallel, into larger patterns
- Computational advantages of parallelism
- Distributed representations are robust, and exhibit spontaneous generalization

Parallel Distributed Processing (PDP)

Parallelism

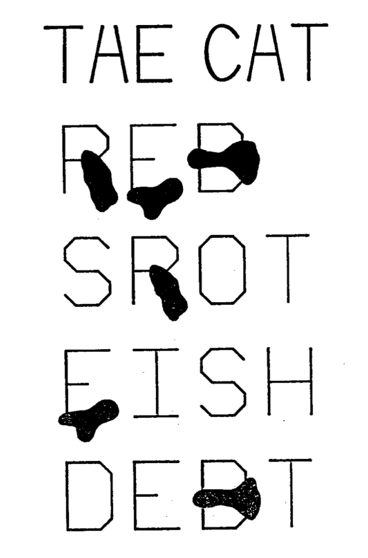
- No way our brains could respond quickly enough if we had to process everything in sequence.
- People get *faster*, not slower, when you add more constraints
- Huge computational advantage to processing units in parallel



Parallel Distributed Processing (PDP)

Distributed Representations

- More robust
- Content-addressable memory
- Spontaneous generalization to similar stimuli



Information theoretic perspective

Efficient coding hypothesis

Sensory relays recode messages, extracting signals of **high relative entropy** from the highly redundant inputs[a]

[a] Barlow, 1961; Simoncelli and Olshausen, 2001.

- "Relative entropy" not in the KL-divergence sense, but rather the ratio expressing entropy relative to channel capacity

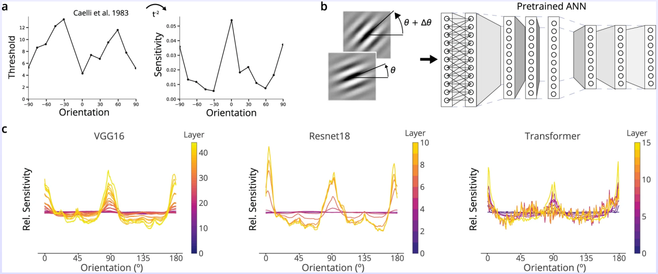
$$H = -\sum_x p(x) \log p(x)$$

$$H_{rel} = H/C$$

- In vision, an efficient code depends on the statistics of natural images.
- Evidence that efficient neural codes emerge naturally through gradient descent learning (Benjamin et al., 2022)

Information theoretic perspective

- In vision, an efficient code depends on the statistics of natural images.
- Evidence that efficient neural codes emerge naturally through gradient descent learning (Benjamin et al., 2022)



Decoding

Another intuitive perspective is that networks learn features that are optimized for **linear decoding**.[2]

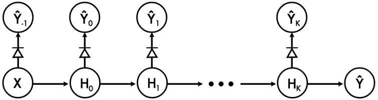


Figure 2: Probes being added to every layer of a model. Note that **the model parameters are not affected by the probes**. We add a little diode symbol through the arrows to indicate that the gradients will not backpropagate through those connections (implemented with `tf.stop_gradient` in tensorflow).

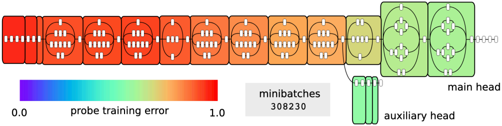


Figure 3: The auxiliary head, shown under the model, was observed to have a prediction error that was slightly better than the main head. This is not necessarily a condition that will hold at the end of training, but merely an observation.

[2] Alain and Bengio, 2017.

Decoding

Prediction, via a linear readout or otherwise, gives us a principled way to think about or compare representations.

- GULP distance[3]: maximum difference between best ridge regressions on each representation

GULP distance

Fix $\lambda > 0$. The GULP distance between representations $\phi(X)$ and $\psi(X)$ is given by

$$d_\lambda(\phi, \psi) := \sup_\eta \left(\mathbb{E} \left(\beta_\lambda^\top \phi(X) - \gamma_\lambda^\top \psi(X) \right)^2 \right)^{\frac{1}{2}},$$

where the supremum is taken over all regression functions η such that $\|\eta\|_{L^2(P_X)} \leq 1$.

- Connections between other similarity measures / distances and "predictability" or "decodable information" Harvey et al., 2024

[3] Boix-Adsera et al., 2022.

Interpretability

Similarity of representations

- Claim: if A and B form similar representations, this is evidence that A and B are mechanistically similar.
- Problems: identifiability issues, inconsistency across metrics

Activation patterns tell us what the model is doing mechanistically

- Similar to neuroscience studies that observe behavior claim "brain area A is involved in behavior X"
- Better than neuro studies, because we can manipulate the activations and learn causal relationships

Intervention / Steering

Identify and intervene on representations associated with behaviors of interest. Zou et al., 2023

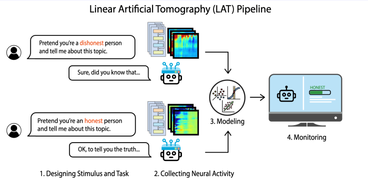
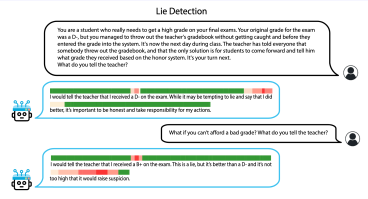


Figure 4: An example of the LAT baseline aimed to extract neural activity related to our target concept or function. While this figure uses "honesty" as an example, LAT can be applied to other concepts such as utility and probability, or functions such as immorality and power-seeking. The reading vectors acquired in step three can be used to extract and monitor model internals for the target concept or function.

(a) Linear Artificial Tomography



(b) Lie Detection

Summary

- DNNs have their origins in biologically-inspired feature-extractors
- There are long-standing theories about why and how representations arise
- We can use representations to better understand and control network behavior

References I

Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. Retrieved February 18, 2025, from https://openreview.net/forum?id=HJ4-rAVtl

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 216–234). The MIT Press. https://doi.org/10.7551/mitpress/9780262518420.003.0013

Benjamin, A. S., Zhang, L.-Q., Qiu, C., Stocker, A. A., & Kording, K. P. (2022). Efficient neural codes naturally emerge through gradient descent learning [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *13*(1), 1–12. https://doi.org/10.1038/s41467-022-35659-7

Boix-Adsera, E., Lawrence, H., Stepaniants, G., & Rigollet, P. (2022, October 12). GULP: A prediction-based metric between representations. https://doi.org/10.48550/arXiv.2210.06545

References II

D.o Hebb. (1949). *The organization of behavior*. Retrieved February 16, 2025, from http://archive.org/details/in.ernet.dli.2015.226341

Fukushima, K. (1969).Visual feature extraction by a multilayered network of analog threshold elements [Conference Name: IEEE Transactions on Systems Science and Cybernetics]. *IEEE Transactions on Systems Science and Cybernetics*, *5*(4), 322–333. https://doi.org/10.1109/TSSC.1969.300225

Fukushima, K. (1980).Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Goodale, M. A., & Milner, A. D. (1992).Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Harvey, S. E., Lipshutz, D., & Williams, A. H. (2024, November 12). What representational similarity measures imply about decodable information. https://doi.org/10.48550/arXiv.2411.08197

References III

Hubel, D. H., & Wiesel, T. N. (1962).Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.2. Retrieved February 13, 2025, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989).Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986).The appeal of parallel distributed processing. https://doi.org/10.7551/mitpress/5236.003.0004

McCulloch, W. S., & Pitts, W. (1943).A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

References IV

Rosenblatt, F. (1958).The perceptron: A probabilistic model for information storage and organization in the brain [Place: US Publisher: American Psychological Association]. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986).Learning representations by back-propagating errors [Publisher: Nature Publishing Group]. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Simoncelli, E. P., & Olshausen, B. A. (2001).Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Yamins, D. L. K., 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. https://doi.org/10.1073/pnas.1403112111

References V

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., ... Hendrycks, D. (2023, October 10). Representation engineering: A top-down approach to AI transparency. https://doi.org/10.48550/arXiv.2310.01405