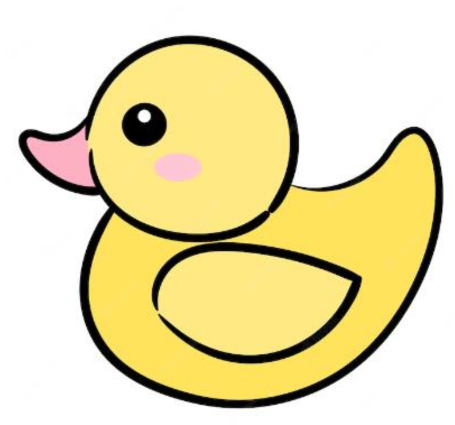
Tensor Product “Variable Binding” and the Representation of Symbolic Structures

Paul Smolensky (1990)
as narrated by
Gerald Penn
University of Toronto
Vector Institute for Artificial Intelligence

Leibniz's Law of the Identity of Indiscernibles (1686)



Filler/role representation: $\beta : S \rightarrow 2^{F \times R}$; $s \mapsto \{(f, r) \mid f/r(s)\}$

Role decomposition: $\mu_{F/R} : F \times R \rightarrow Pred(S)$; $(f, r) \mapsto f/r$.

Leibniz's Law of the Identity of Indiscernibles (1686)

- But these are partial functions (r is the function, not f)
- "Faithful:" 1-to-1, stay away from 0 vector.
- Superposition: conjunction as addition
  - When can we recover the original conjuncts?
  - What about when multiple conjunctions are stored in/learned by the same network?
  - 2 different answers provided:
    - orthonormality (Hebbian learning),
    - linear independence (Widrow-Hoff learning).

“Tensors”

- Independently encode fillers and roles in vector spaces:

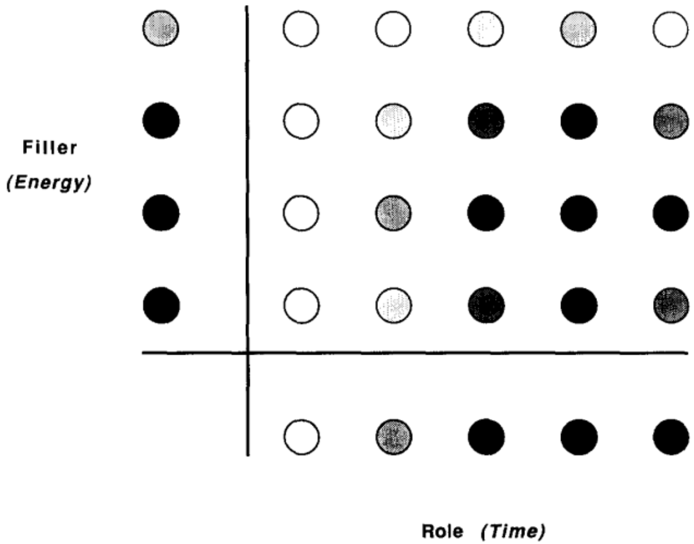
$$\Psi_F : F \to V_F \,, \qquad \Psi_R : R \to V_R$$

- ... then combine with cross-product:

$$\Psi_b : \{ f/r \mid f \in F, r \in R \} \to V_F \otimes V_R \,; \qquad f/r \mapsto \Psi_F(f) \otimes \Psi_R(r) \,.$$

- Matrix → Tensor?
  - A matrix is a kind of tensor, and he will use tensors of higher rank later,
  - …but not all rank-2 tensors are matrices.
- Cross-product → Tensor?
  - More truth to this than perhaps Smolensky is aware

Example: Spectograms

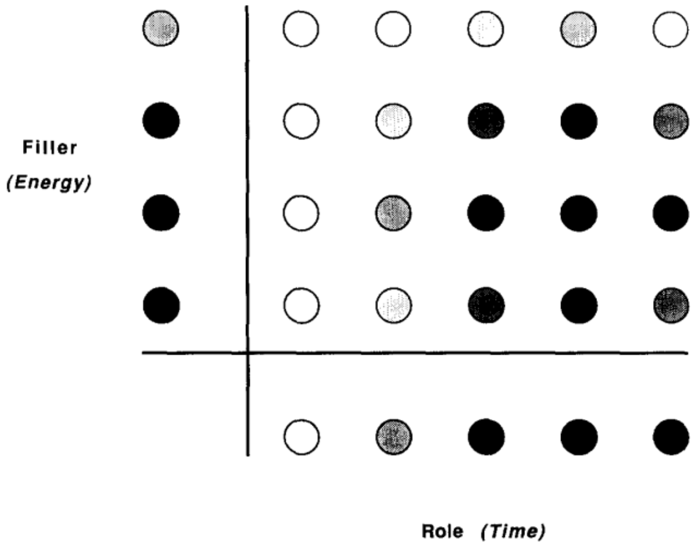


- Very discrete, but not as oversimplistic as you may think: sampling.
- See also “continuous strings,” p. 191.

Nota bene!

- p. 200: “primary purpose of the tensor product rep is not to serve as an apparatus for filler/role associations ...rather to provide pattern of activity representing a structured object which can then be used [as a proxy for the whole object during processing].”
- Smolensky (1990) will turn to this later in the paper, but in a way that seems to imply it’s only important because of the existence of recursion in language.
- Unitary matrices aren’t popular just because they stand in a 1-to-1 relation with complex numbers.

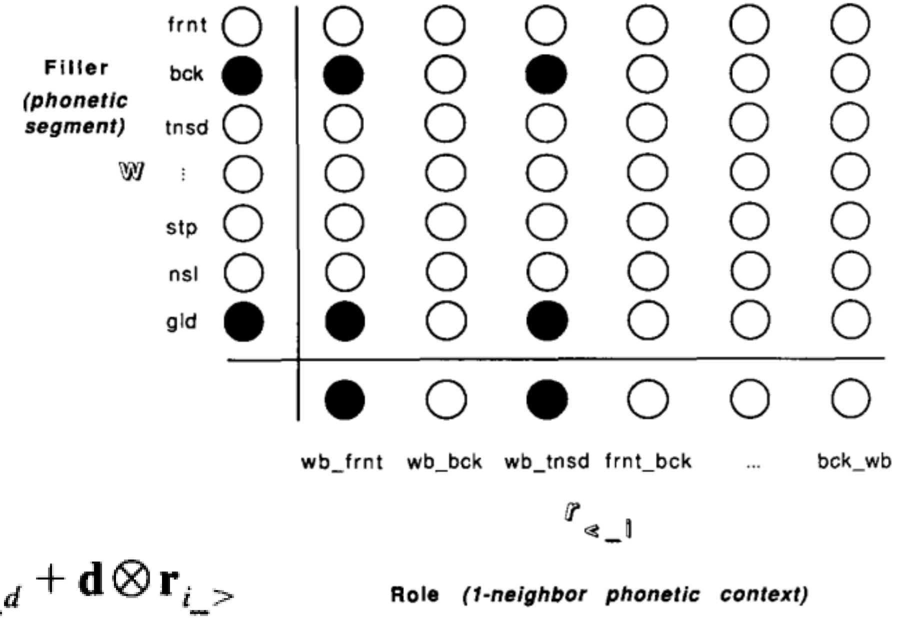
Example: Phonetic string decompositions



- Lots of possibilities, most of them obvious now, but big deal back then.
  - “Local” – 1-hot encodings
  - “Distributed” – over pools(!) of units
- See also “continuous strings,” p. 191.

Example: Phonetic string decompositions

- Lots of possibilities, most of them obvious now, but big deal back then.
  - “Local” – 1-hot encodings
  - “Distributed” – over pools(!) of units
- The winner:



$$\mathbf{w}\otimes\mathbf{r}_{<\_j}+\mathbf{i}\otimes\mathbf{r}_{w\_d}+\mathbf{d}\otimes\mathbf{r}_{i\_>}$$

O tempora, O mores…

- p. 185: “…the answer depends on dynamics driving the connectionist network, not solely on the representations themselves“
  - vs. tacit assumption in much of deep learning today that CL is now purely a science of representation: no algorithms, no algebra.
- p. 193: “Here it is not the job of the network to set up tensor product representations: ... the modeler must convert the symbolic inputs and outputs to their vector representations"
  - not anymore: LLMs, end-to-end architectures, etc.

The Big Theorem

Theorem 3.3. *Suppose the self-addressing procedure is used to unbind roles. If the role vectors are all orthogonal, the correct filler pattern will be generated, apart from an overall magnitude factor. Otherwise, the pattern generated will be a weighted superposition of the pattern of the correct filler,* $\mathbf{f}_i$, *and all the other fillers,* $\{\mathbf{f}_j\}_{j \neq i}$. *In this superposition, the weight of each erroneous pattern* $\mathbf{f}_j$ *relative to the correct pattern* $\mathbf{f}_i$, *the* intrusion of role j into role i, *is*

$$\frac{\mathbf{r}_i \cdot \mathbf{r}_j}{\|\mathbf{r}_i\|^2} = \cos\theta_{ji} \frac{\|\mathbf{r}_j\|}{\|\mathbf{r}_i\|}$$

where θ_{ji} *is the angle between the vectors* $\mathbf{r}_j$ *and* $\mathbf{r}_i$.

- Strong connections to Hebbian learning of associations (discussed).
- Replace Hebb with Widrow-Hoff? Then linear independence is key.
- Strong connections to algebraic topology (not discussed).

When Things Go Wrong

Definition 3.4. Let F/R be a role decomposition of S. A connectionist representation Ψ of S has *unbounded sensitivity* with respect to F/R if for arbitrarily large n,

$$\Psi\left(\bigwedge_{i=1}^{n} f_i/r_i\right)$$

varies as f_i varies, for each $i = 1, 2, \ldots, n$.

If for sufficiently large n the representation of structures containing n filler/role bindings is not faithful, then Ψ *saturates*.

If Ψ saturates and has unbounded sensitivity then Ψ possesses *graceful saturation*.

- Saturation bounds (Thm 3.5, pp. 188-9): number of bindings that can be stored before magnitude of intrusion = magnitude of correct pattern is $O(\sqrt{N})$, N = dimensionality of role vectors.

Generalizing to Continuous Case

Definition 3.6. Let F/R be a role decomposition of S, not necessarily finite, and let $\mathrm{d}\mu(r)$ be a measure on R. Let $\mathrm{supp}_R(s)$ be the subset of R containing roles which are bound in s, and suppose F/R has single-valued roles. Suppose given the connectionist representations

$$\Psi_F : F \to V_F\,; \qquad f \mapsto \mathbf{f}\,,$$

$$\Psi_R : R \to V_R\,; \qquad r \mapsto \mathbf{r}$$

and assume these functions are measurable with respect to $\mathrm{d}\mu(r)$. Then the corresponding tensor product representation of S is

$$\Psi_{F/R}(s) = \int_{\mathrm{supp}_R(s)} \mathbf{f}(r) \otimes \mathbf{r}\, \mathrm{d}\mu(r)\,.$$

$\Psi_{F/R}(s)$ is defined only for those s for which the integral is well-defined: $\mathrm{supp}_R(s)$ must be a measurable set and the integral must converge.

- Both the roles (“continuous strings”) and the fillers (e.g. Gaussians) can be continuous.

Recursion

- Filler recursion (iterative application to more arguments, e.g., list): increase dimensionality.
- Role recursion (higher-order role, application of which results in lower-order role): increase rank of tensor.
- Discussion here is foreshadows fascination 10 years later with *monads*:

$$\eta : \alpha \to \mathbb{M}\alpha, \qquad \star : \mathbb{M}\alpha \to (\alpha \to \mathbb{M}\beta) \to \mathbb{M}\beta.$$

$$\begin{aligned}
\eta(a) \star k &= k(a) && \forall a : \alpha,\ k : \alpha \to \mathbb{M}\beta, \\
m \star \eta &= m && \forall m : \mathbb{M}\alpha, \\
(m \star k) \star l &= m \star \big(\lambda v.\, k(v) \star l\big) && \forall m : \mathbb{M}\alpha,\ k : \alpha \to \mathbb{M}\beta,\ l : \beta \to \mathbb{M}\gamma.
\end{aligned}$$

Annihilators

Definition 3.15. Let F/R be a role decomposition of S and let $k \mapsto s^{(k)}$ be a sequence of elements in S. An *annihilator* of $k \mapsto s^{(k)}$ with respect to R/F is a sequence of real numbers $k \mapsto \alpha^{(k)}$, not all zero, such that, for all fillers $f \in F$, and all roles $r \in R$,

$$\sum_{k\,:\,f/r \in \beta(s^{(k)})} \alpha^{(k)} = 0\,.$$

For example, consider the sequence of strings (ax, bx, ay, by). With respect to the positional role decomposition, this has annihilator $(+1, -1, -1, +1)$, since for each filler/role binding in $\{a/r_1, b/r_1, x/r_2, y/r_2\}$, the corresponding annihiulator elements are $\{+1, -1\}$, which sum to zero.

- These are an important caveat to the linear independence result for Widrow-Hoff learning.
- Again, algebraic topology provides the tools for thinking about this systematically.

Implementation!

- TPPS: tensor product production system
- Dolan and Dyer (1988-9)
- Source code available?