Representations in Simple Recurrent Networks Which are Always Compositional

David Landy (dlandy@indiana.edu)
Departments of Computer Science and Cognitive Science, Indiana University
107 S. Indiana Ave., Bloomington, IN 47405-7000

In classical cognitive models, representations of inputs are deliberately built into the operational structure by a model's designers. Network systems by contrast usually automatically construct responses following some generic learning scheme, and consequently lack overt representations altogether. Instead, the system's representations are read off the system according to a chosen analytical methodology. The performance of such models is therefore independent of how their representations are labeled.

Simple recurrent networks (SRNs) are among the most successful network models of cognition (Elman, 1990, 1995). These networks are often taken to represent inputs in the values of their hidden layer nodes, which can be analyzed using principal component analysis or hierarchical clustering. Under this interpretation, representations in networks are context-sensitive, static, and non-compositional. Significantly different properties result from taking as the representation of a sequence the function which that sequence causes the network to compute.

Consider a typical SRN with input weights W_{in}, output weights W_{out}, and recurrent connections in the hidden layer with weight matrix C, and call the vector of weights in the hidden layer H. Let S denote the closure of the set of legal inputs to the network under concatenation, so that S contains all legal sequences (and also an empty input, ε). Call the set of possible output vectors O, and call the function which maps input sequences to output vectors $i:S \rightarrow O$, so that $i(s)=o$ exactly when o is the output resulting from running sequence s through the network. Consider the following function:

$$r'(s,h) = \begin{cases} h & \text{if } s = \varepsilon \\ r'(u, Sigmoid(C \cdot h + W_{in} \cdot t) & \text{if } s = t.u \end{cases}$$

r' can be interpreted as the function which computes, for an initial value on the hidden layer, h, the value on the hidden layer which results after processing input sequence s. Define the family of functions which results from currying r' over s: $r_s(\vec{h}) = r'(s,\vec{h})$, $r_s \in R$. Then R is the representation scheme of the SRN. $i(s)$ can be easily reconstructed from r_s.

A straightforward homomorphism can be constructed between concatenation over S and function composition in R, making the representations of any SRN classically compositional, regardless of the prior training of the network (see Fodor & Lepore, 2002; Zadrozny, 1994). This analysis also reveals a limited form of inherent systematicity, in that the same representation function, and hence the same causal mechanism, is employed in processing a particular lexeme or sequence regardless of the context in which it appears (see Davies, 1991).

If the computation specified by r_s picks out a type of representation, then any particular application of that function can be taken to be a token. The computation performed is independent of its context, but the specific hidden layer value which results will not be. Therefore, tokens of computations can be picked out by specifying the input/output pair (where both input and output are hidden layer values) which that application involved. The method of hierarchical clustering which is so useful in analyzing hidden layer values can then be performed on this pair, and so this technique can be applied essentially unchanged. Additionally, the extra information stored in the source values allows the method to be applied to sequences as well as single inputs.

Since these representations are the system's disposition to respond to a particular lexeme, rather than the residue of state information which results from that response, these representations are active processes rather than static data structures. Since the important

Therefore, representations capture all of the knowledge which is involved in generating the internal state of the network.

Because the representation scheme given here appropriately encapsulates an SRN's knowledge and presents representations as dynamic processes rather than static structures, it is intuitively appealing as a model for how SRNs represent. Inasmuch as it is appealing, SRNs represent compound phrases compositionally and context-independently, which implies that these properties may not account for some of the interesting properties with which they have been credited (Fodor & Lepore, 2002).

References

Davies, P. 1991. Concepts, connectionism, and the language of thought. In William Ramsey, Stephen Stich, David Rumelhart, eds., *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science* 14:179-211.

Elman, J. L. 1995. Language as a dynamical system. In Port, R. F. and Van Gelder, T., eds., *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, 195-223.

Fodor, J., and Lepore, E. 2002. Why meaning (probably) isn't conceptual role. In *The Compositionality Papers*. Oxford University Press.

Zadrozny, W. 1994. From compositional to systemantic semantics. *Linguistics and Philosophy* 17:329-342.