Modelling Language Acquisition at Multiple Temporal Scales

Steve R. Howell (showell@hypatia.psychology.mcmaster.ca)
Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada

Suzanna Becker (becker@mcmaster.ca)
Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada

The problem of incorporating time in a neural network is an important one. Networks with feedback, such as Simple Recurrent Networks (SRNs) (Elman, 1990) have been argued to represent time more realistically through its effects on the processing of input, compared to standard feedforward networks. In effect, SRNs' context units act as a memory, which incorporates a "smeared-out" representation of the network's internal states over time.

A problem does exist with this representation, however, especially for complex domains like that of language. The nature of argument agreement, embeddings and similar phenomena means that the SRN must be able to represent important past states (such as head noun for verb agreement) in spite of the declining effects of past context. While most word inputs will be related most strongly to words co-occurring close by in the input stream, verb agreement, for example, is largely determined by its corresponding noun, even in long, multiply-embedded sentences. SRN models should therefore be able to preserve representations of vital early structure for later use, in spite of the generally appropriate decline of short-term context. This issue has been addressed in an architectural fashion by others (Weckerly and Elman, 1992), but can perhaps be addressed more generally by allowing for more than one duration of context in an SRN's operation.

It is possible to apply the concept of hysteresis to the SRN's context units. That is, the update from the hidden units to the context units may be other than the usual 1-to-1 copying; the context units may also incorporate self-recurrent connections of varying strengths. In particular, we have been experimenting with SRN's using the hysteresis function suggested by Wermter, Arevian, and Panchev (1999) on the self-recurrent connections:

$$Context_i(t+1) = (1-Hy)*Hidden_i(t) + Hy*Context_i(t)$$

We have conducted initial experiments using a test corpus derived from the original simplified test corpus used by Elman (1990). Our version differs from the original in that it includes not only consonant to vowel relations, but also word-to-word relations. That is, some of the consonant-vowel combinations (words) can only occur immediately following others.

Thus in addition to the network needing to learn, for example, that u's only come after G's or U's (Guuu), it must also learn that Guuu only comes after Da. It is in this capacity that the hysteresis parameter should most come into play, for it specifies, in effect, the duration of retention of the states of the context units. For short term letter to letter relations, small to zero hysteresis values should be adequate, as demonstrated originally by Elman's success. In that experiment, the network error declined consistently within a word, but jumped at word boundaries, representing the fact that word distribution was random in that corpus. In our experiments, manipulating the hysteresis parameters was expected to bias the network in favour of either short or long term relationships. Also, simulated annealing of the learning rate, another technique not typically used with SRNs, is used in both control and experimental networks. In pilot work this feature smoothed oscillations in the gradient descent of error.

The initial results of a number of simulation runs from different random initial conditions indicate that small hysteresis values (of other than 0) are indeed an advantage in learning this prediction task, with error per epoch declining noticeably, though not exceptionally, faster with $0.2 > Hys > 0.1$. Presumably this modest net gain is actually composed of both a larger gain for word-to-word relationships and a small decline for letter-to-letter prediction. Explorations of the exact nature of this advantage are underway, as is investigation of the best range of hysteresis parameters for various language tasks.

With the ability to change the hysteresis of context layers, it becomes useful to incorporate multiple hidden layers into an SRN (Wermter, 1999), with layers having a different 'span' of context via different hysteresis settings. We also describe a model with multiple hidden layers that is being applied to more complex language corpora, and is designed to be able to learn at multiple time scales simultaneously, by capturing longer-range temporal structure in progressively higher layers.

References

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

Weckerly, J., & Elman, J.L. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Wermter, S., Arevian, G. & Panchev, C. (1999). Recurrent Neural Network Learning for Text Routing, *Proceedings of the Ninth International Conference on Artificial Neural Networks*, 2, 470-475.