Visual Cues in Connectionist Models of Language Acquisition

Bo Pedersen (bop@cornell.edu)
Department of Psychology, Cornell University
Uris Hall, Ithaca, NY 14853 USA

In the debate about the learnability of language it has been argued that children can use bootstrapping of various information to overcome the otherwise daunting problem of language learning (Christiansen & Dale, 2001). Simulations suggest that the integration of distributional, phonological, and prosodic cues can promote better, faster and more uniform learning of syntactic structure in simple recurrent networks (SRN's)

Recent studies have shown that similar results can be obtained with sensorimotor cues (Howell, Jankowicz & Becker, 2005). The addition of semantic cues like hard, soft, sweet, indoor, in-kitchen, does-fly, pleasant, and man-made, can promote syntactic learning in SRN's.

These categories are very abstract and must to some degree be learned prior to syntactic learning for the argument to work. And thus the methodology faces another type of bootstrapping problem. A possible way to overcome this problem is to go to the source itself: the raw input from our surrounding sensorimotor world, and explore radically unsupervised language learning based on recordings of co-occurring language and sensorimotor data.

One rich source of such data is vision. The problem of using vision is that it has its own bootstrapping problems in the dichotomy of object and scene. This problem can be solved in various ways. One way is to relax this dichotomy and not make a strict structural analysis of the scene, but just look at the "gist" of a scene. The gist can then in turn give probabilistic information about scene and object: If the gist is forest-like we might be seeing a tree and if it is tree-like we might be seeing a forest (Torralba, Murphy, Freeman & Rubin, 2003).

For a language learning task the gist of a scene might be enough. Children probably don't need full object recognition to learn language. They will use any kind of statistics they can get from the outside world.

Method

The gist cues for the present simulation are 3430 feature vectors of length 80, extracted with the steerable pyramid method from a set of real world images from 20 different scenes collected with a head mounted webcam (available on cs.ubc.ca/~murphyk/Vision/placeRecognition.html)

Three SRN's were trained on 4000 random sentences on a grammar similar to Christiansen & Dale's (2001), but with only 19 tokens since the number of scenes is only 20. The first network has an input layer consisting of a localist representation of a word (19 nodes) and a visual gist (80 nodes) and a hidden layer consisting of 160 nodes and an output layer with the next word to be predicted. To each input word we attach a random gist belonging to that word's scene type. The second (control) network has a random gist from *any* scene type and the third (control) network has no gists at all and is just the pure language learning task.

Results

On each network 100 simulations with 100 epochs each were run with new random word/scene couplings every time and the networks were tested on a 1000 novel word/scene couplings. The gist network performed significantly better ($p<.001$, t-test on MSE on the last 20 epochs) than the two control networks with an MSE of .03 on the training sentences and .04 on novel sentences. The random gist control performed at.06/.1 and the small no-gist network at .04/.05.

The coupling of words and gists are admittedly a bit arbitrary and the experiment assumes that the visual system can provide word-sized gists from full complex scenes and relate these directly to words. In the future we would like to make more complex simulations that directly relate whole scenes to whole sentences, if we can find such data. But we find that the current simulations demonstrate that it might be possible to pursue the direction set out by Howell, Jankowicz & Becker (2005) in a more data-driven fashion, using data directly from our sensorimotor world rather than abstract representations.

References

Christiansen, M.H. & Dale, R.A.C. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society* (pp. 220-225). Mahwah, NJ: Lawrence Erlbaum

Howell, S. R. & Jankowicz, D. & Becker, S. (2005). A Model of Grounded Language Acquisition: Sensorimotor Features Improve Lexical and Grammatical Learning. *Journal of Memory and Language* (pp. 258-276). Orlando, FL: Elsevier.

Torralba, A., Murphy, K. P., Freeman W. T. & Rubin, M. (2003) Context-based vision system for place and object recognition. *Proceedings. Ninth IEEE International Conference on Computer Vision*, (pp.). Washington, DC: IEEE Computer Society