Teaching Newton to speak: Using CSAR services to speed the training of a neural network that models human language processing

Stephen Welbourne and Matthew Lambon Ralph
University of Manchester

Introduction

Neural networks are now well established tools in the study of language processes (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Harm & Seidenberg, 2004; Plaut, 1996; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989; Welbourne & Lambon Ralph, 2005). These networks are attractive as models because they process information in a similar way to the human brain; using a large number of simple information processing units, in parallel, to map representations across domains. In the case of the brain these processing units are neurons whereas in the models they are artificial neurone-like constructs, built from code that runs on a serial processor (usually a PC). In both cases learning occurs as a result of modification to the weights (synaptic strengths) of the connections between units, with these modifications occurring slowly over a large number of repeated trials.

The main limiting factor in the use of these models is the computational resource that is required to train them. Unlike the brain, these models cannot truly process information in a parallel manner, but have to simulate parallelism by cycling through the units serially. In addition, the kind of language tasks that are interesting to model tend to require training on large corpuses of words, typically thousands of items. As a result it is not unusual for these models to require weeks or even months of processing.

The purpose of this Class 3 project was to test the feasibility of using parallel supercomputers to significantly reduce the processing time required for this kind of model. Ultimately we would like to model speech, verbal comprehension and reading behaviours simultaneously, within the same generalised language model. However, for the purposes of this simulation, we elected to concentrate solely on the mapping from meaning to phonology (speech).

Simulation Details

The training corpus consisted of 2998 monosyllabic words with phonological representations taken from Plaut et al. (1996). The semantic representations were constructed by generating unique random binary vectors of length 100 with an average of 20 units set to 1 and 80 set to 0. This ensured that we preserved two important features of human semantics: firstly, that semantic representations are relatively sparse, and secondly, that the mapping between semantics and phonology is not in any way systematic.

Figure 1 shows the architecture of the recurrent network that was used for these simulations with semantic and phonological layers connected by hidden layers consisting of 1500 units. Where layers of units are shown as connected it was always the case that every unit in the sending layer was connected to every unit in the receiving layer. Activation functions for the units were logistic with time integrated inputs. The network was trained using standard backpropogation through time with a learning rate of 0.05 and momentum of 0.9, applied only when the gradient of the error slope was less than 1.

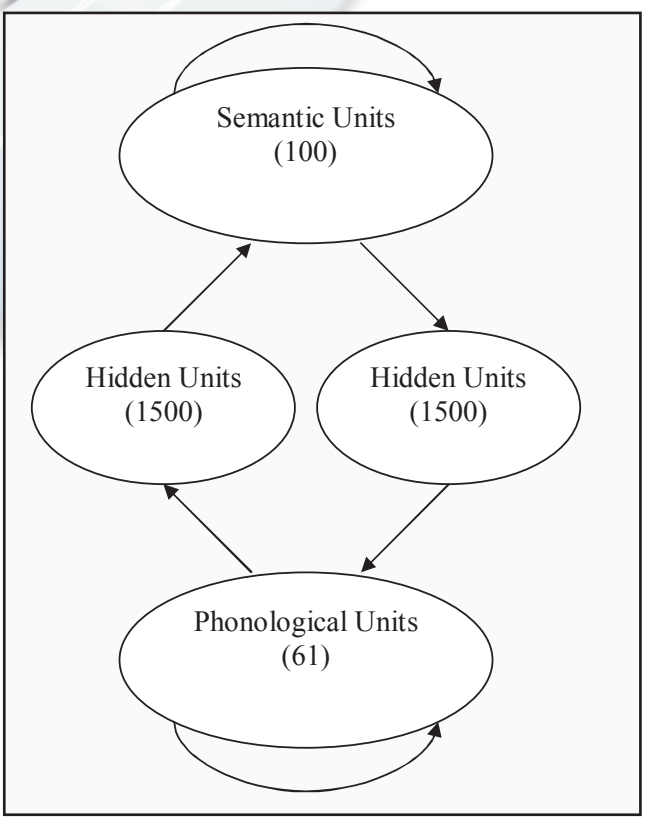

Figure 1: Network Architecture.

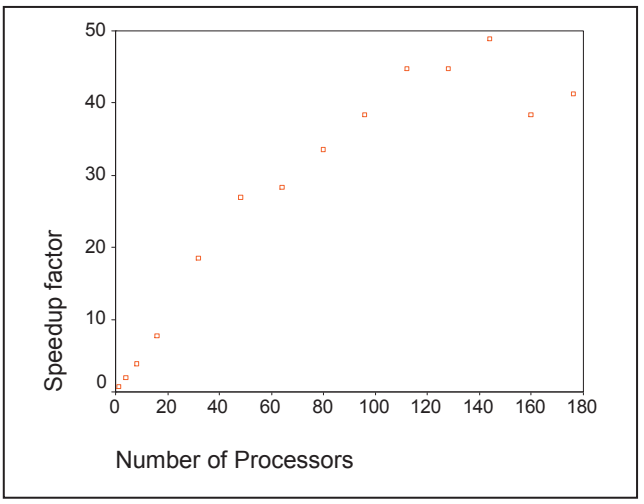

Figure 2: Showing how speed of training scales with number of processors compared to a desktop PC.

Parallelisation Method

In this task there are two obvious approaches to parallelisation. Either one can parcel out the units between processors, or one can replicate the entire network on each processor and divide the training batch by processor. We elected to adopt the latter approach and, with the extensive help of the CSAR support team, we adapted our existing network simulator code to run on Newton.

Results

For the purposes of this feasibility study, we were not interested in how well the network could perform the task, but merely in how long it took to accumulate the weight updates for one pass of the entire training corpus (1 epoch of training). In, particular we were interested in how the speed of training would scale with the number of processors. Accordingly, we ran trials of 2 epochs of training over increasing numbers of processors up to a maximum of 176. Figure 2 shows the results of this investigation. For convenience the y scale is expressed in multiples of single processor speeds (measured on a standalone Pentium 4 3.2Ghz PC running windows XP). When using only one processor Newton actually runs the code more slowly than on a standalone PC (speedup=0.76). However, the speed of processing scales reasonably linearly all the way up to 100 processors (speedup≈40). After this processing speed continues to improve slightly up to about 140 processors. Beyond that adding extra processors actually reduces processing speed.

Discussion

This project set out to test the feasibility of using supercomputing services to speed the training of neural networks modelling linguistic processes. Using a typical network setup, modelling the mapping from semantics to phonology, we have demonstrated that speedup factors in excess of 40 are achievable. Time constraints prevented us from conducting further empirical investigations; it would be interesting to know how the network parameters (number of units and batch size) would affect the scaling performance. Nevertheless, we have clearly shown that this approach has considerable potential to reduce the time required to train these kinds of networks.

References

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. Psychological Review, 104(4), 801-838.

Harm, M.W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. Psychological Review, 111(3), 662-720.

Plaut, D.C. (1996). Relearning after damage in connectionist networks: toward a theory of rehabilitation. Brain And Language, 52(1), 25-82.

Plaut, D.C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains. Psychological Review, 103(1), 56-115.

Seidenberg, M. S., & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. Psychological Review, 96(4), 523-568.

Welbourne, S. R., & Lambon Ralph, M. A. (2005). Exploring the impact of plasticity-related recovery after brain damage in a connectionist model of single-word reading. Cognitive, Affective & Behavioral Neuroscience, 5(1), 77-92.