A Comparison of Elman and Echo State Networks

Elliott M. Forney
Colorado State University
Computer Science Department
idfah@cs.colostate.edu

Charles W. Anderson
Colorado State University
Computer Science Department
anderson@cs.colostate.edu

Poster Abstract

Artificial Neural Networks (ANN) are powerful, trainable approximation structures. ANN's are composed of a number of simple computational units, or neurons, with weighted interconnections. Training is typically performed by adjusting the connection weights between these neurons in a way that minimizes the error produced with respect to a given mapping. Recurrent Artificial Neural Networks (RNN) are a special class of ANN that contain delayed feedback connections. These feedback connections give RNN's an intrinsic state and the ability to approximate mappings that require memory. There are numerous RNN architectures and corresponding optimization algorithms, all with different characteristics. Here, we compare two different types of RNN by analyzing their performance on several benchmark problems.

Elman's Simple Recurrent Networks (SRN) consist of two layers. In the first, or hidden, layer, the computation performed by the neurons is a hyperbolic tangent. All neurons in the hidden layer are densely connected to the network inputs and have feedback connections to all other neurons in the hidden layer with a delay of a single timestep. The computation performed by the neurons in the second, or visible, layer is a simple linear combination. Neurons in the visible layer are densely connected to the hidden layer and contain no feedback connections. SRN's are typically trained by unfolding the network a number of steps back through time and then removing all feedback connections. The error gradient of the original network can then be approximated using this acyclic network and a standard gradient descent algorithm can be applied to minimize training error. SRN's were originally proposed by Jeff Elman in 1990 and a theoretical result has since shown that SRN's can approximate any finite state machine with arbitrary precision, given enough hidden units and the proper weight values. However, real-world applications that utilize SRN's are relatively rare due to the difficulty of avoiding local optima during training.

ESN's have a similar layout to SRN's in that they have a single hidden layer with a hyperbolic tangent transfer function and a visible layer with a linear transfer function. The hidden units in an ESN also have feedback connections with a delay of a single timestep. Unlike SRN's, however, the hidden layer of an ESN is relatively large, sparsely connected and remains untrained. Instead of adjusting the weights of the hidden layer, the weights are chosen to have the Echo State Property. That is, the output of the hidden layer will asymptotically converge to the same state given the same input sequence, regardless of initial conditions. Training the visible layer then becomes a simple linear regression problem. ESN's were originally proposed by Herbert Jaeger in 2002 and are one of several types of approximation structures that depend on large untrained "reservoirs." ESN's have received significant attention in recent neural network literature.

We compare SRN's and ESN's by applying both approaches to several benchmark problems. The first is known as the temporal *XOR* problem. In the temporal *XOR* problem, a sequence of bits are fed into the network which is then trained to output the exclusive *OR* of the bits that were seen a number of steps, T and $T-1$, back in time. The memory requirement of the network can be increased in the temporal *XOR* problem by increasing the value of T. The temporal *XOR* problem is an important benchmark because it is both non-linear and contains a temporal dependency. The second benchmark is a simple time series forecasting problem in which each RNN is trained to forecast a sinusoidal function a number of steps ahead in time. The difficulty of this problem can also be increased by requiring the network to forecast the signal further ahead in time. The final benchmark is a more difficult time-series forecasting problem using a laser-generated dataset borrowed from The Santa Fe Time Series Competition.

Preliminary results suggest that ESN's are able to achieve significantly longer term memory in the temporal *XOR* problem and typically outperform SRN's in the forecasting problems. However, ESN's can exhibit numerical instability problems with some choices of initial weights. Additionally, it may be more difficult to control overfitting when using ESN's. Although further investigation involving a wider range of problems will undoubtedly be required, it appears that the recent enthusiasm surrounding ESN's may be well founded.