Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005

Reservoir Riddles: Suggestions for Echo State Network Research (Extended Abstract)

Herbert Jaeger
International University Bremen
Bremen, Germany
E-mail: h.jaeger@iu-bremen.de

Abstract— **Echo state networks (ESNs) offer a simple learning algorithm for dynamical systems. It works by training linear readout neurons that combine the signals from a random, fixed, excitable "dynamical reservoir" network. Often the method works beautifully, sometimes it works poorly – and we do not really understand why. This contribution discusses phenomena related to poor learning performance and suggests research directions. The common theme is to understand the reservoir dynamics in terms of a dynamical representation of the task's input signals.**

I. Reservoir riddles ...

Echo state networks (ESNs), as well as the closely related "liquid state machines" (LSM) (1), present a recurrent neural network (RNN) learning architecture which is characterized by

- a large, randomly connected, recurrent "reservoir" network that is passively excited by the task's input signal, and
- trainable readout neurons that combine the desired output from the excited reservoir state.

Training an ESN on a supervised learning task boils down to compute the output weights. From a computational perspective this is just a linear regression, for which numerous batch and adaptive online algorithms are available. This simple method yields models that in many engineering tasks surpass in accuracy other modelling methods (2). The ESN/LSM principle – combine a target signal from random, dynamic input variations – may also be effective in biological brains (3) (4).

It is intutively clear that reservoir properties are of great importance for the learning performance.

A basic, necessary property is the *echo state property*: for the ESN learning principle to work, the reservoir must asymptotically forget its input history. A necessary and a sufficient algebraic condition on the reservoir weight matrix are known, which ensure the echo state property (5). Furthermore, a number of heuristic tuning strategies for the three most important global control parameters (network size, spectral radius of reservoir weight matrix, scaling of input) have been described (6). All in all, this body of knowledge renders ESNs applicable in daily practice.

However, this state of the art is clearly immature. Here is a choice of unresolved issues:

- If the reservoir is fixed up to a global weight scaling (for input weights and internal reservoir weights), different tasks require different global scaling parameters for optimal performance. It is not quite clear however which properties of a task have what influence on these scalings.
- It is sometimes observed that the correlation matrix of the activation signals of the reservoir has an eigenvalue spread in the order of 1E12 or even higher. This is typically accompanied by very large learnt output weights (order of 1E8 is easily reached). This is a condition that should better be avoided because

  1) a large EV spread makes it impossible to use the low-cost LSM online learning algorithm,
  2) large output weights imply a lack of generalization capabilities (the trained network will behave very different if input characteristics change but slightly away from the training data),
  3) large output weights require high-precision representations of reservoir state, rendering such networks unsuitable for analog (cheap and fast) VLSI implementations,
  4) in networks featuring output feedback (which implement NARMA filters) large output weights are indicative of marginal or lacking stability,
  5) very large weights which at the same time require a high precision are biologically implausible.

  Adding noise to the reservoir during training very much reduces the EV spread and improves stability in networks with output feedback, but it impairs model accuracy. It is not understood which types of tasks induce a large EV spread and why.
- With standard gradient-descent training methods, or with evolutionary optimization methods, sometimes for a given task a very small (less than 10 or even less than 5 units), yet quite precise RNN model is found. However, reasonably accurate ESNs for these tasks need 100 units or more. It is not understood how task properties relate to the required ESN size.
- On some tasks (communicated to the author by Danil Prokhorov) of the "learning with fixed weights" (aka "metalearning") type, the ESN approach yields results that are quite inferior to models learnt by expertly applying the EKF learning method (7). Apparently the

0-7803-9048-2/05/$20.00 ©2005 IEEE

randomly created ESN reservoirs are "almost surely unsuited" for these particular tasks. Why?

All of these difficulties point in the same direction: the connection between task specifics (dynamical properties of the input and output signal) and properties of the induced reservoir dynamics is not well understood.

II. ... and Reservoir Research.

Here is a list of research questions that in the author's view mark the route to further progress:

- How can one characterize that a given reservoir is suited for a particular task? Currently, the only (rather tautological) answer is: a reservoir is suited if it yields accurate models. A necessary but not sufficient condition for "suitedness" is a low EV spread. If we were dealing with linear systems, an obvious candidate for a "good" reservoir would be one where the unit signals are decorrelated. This would mean, in the perspective of linear systems, that they represent the main orthogonal components of the driving (= input) signal. However, when it comes to nonlinear systems, these metaphors quickly lose their value.
- How can one adapt the reservoir in an unsupervised fashion to the task's type of data? What would be the target quantities that one would wish to optimize by such an unsupervised training? Candidates that come to mind are the EV spread (minimize it – but how?), or a pairwise decorrelation of reservoir signals (the author tried to achieve this with anti-Hebbian learning and failed...), or entropy of reservoir state distribution (maximize it)... all of which aim at making the individual reservoir units as mutually different as possible in some information-theoretic sense.
- What is the role of topological organization of reservoirs? So far, the author mostly worked with randomly and sparsely connected reservoirs that had no "retinal" or otherwise locally homogeneous topology. In contrast, biological (vertebrate) brains clearly exploit spatial segregation of dynamics to realize information-rich dynamical input representations.
- The problem of very large output weights can be very much alleviated if for a given single output channel y not only a single set of output weights is learnt, but instead several of them, plus a switching mechanism (in the spirit of mixtures of experts) that chooses and activates the most appropriate set of output weights according to the current dynamic context. With this method, dramatic jumps in model quality (or equivalently, dramatic reduction of reservoir size) have been achieved in some cases (to be presented at the IJCNN talk).
- If the stake is to obtain small-sized ESNs, employ evolutionary optimization algorithms to pre-adapt the reservoir to tasks of a desired class (first investigations of this kind in (8)). Theoretical problem: what characterizes a canonical *class* of problems such that problems from that class can be learnt with a single ESN?
- Looking at biological brains, shouldn't we expect that the powers of quickly adaptive information processing arise from numerous dynamic feature extractors (carefully optimized by evolution) that transform the sensor input (and importantly, its history) into a wealth of maximally independent signals? The work about slow feature analysis (9) is very inspiring in this respect.

All in all, my personal view at the moment is that ESN/LSM reveal a nice "readout and learn" trick, but the real wonders of learning and adpatation lie in the riddles of features and representations. The true value of ESN/LSMs may lie not in their raw learning performance that we currently experience – naively amazed – but rather in that they give us novel means to characterize and evaluate the quality of internal representations of dynamic sensor input. Namely, a representation is "good" if it enables fast, robust *learning* of desired output signals. The contribution of ESNs to this eternal question may be that ESNs disentangle the representation (in the reservoir) from learning (of the output weights), which in previous RNN learning schemes were tied together.

References

[1] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Computation*, vol. 14, no. 11, pp. 2531–2560, 2002. [Online]. Available: http://www.lsm.tugraz.at/papers/lsm-nc-130.pdf

[2] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, pp. 78–80, 2004.

[3] G. B. Stanley, "Recursive stimulus reconstruction algorithms for real-time implementation in neural ensembles," *Neurocomputing*, vol. 38-40, pp. 1703–1704, 2001.

[4] W. M. Kistler and C. I. De Zeeuw, "Time windows and reverberating loops: A reverse-engineering approach to cerebellar function," *The Cerebellum*, 2002, in press, http://www.eur.nl/fgg/neuro/research/Kistler/reverse_engineering.pdf.

[5] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks," GMD - German National Research Institute for Computer Science, GMD Report 148, 2001, http://www.faculty.iu-bremen.de/hjaeger/pubs/EchoStatesTechRep.pdf.

[6] ——, "Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach," Fraunhofer Institute AIS, http://www.faculty.iu-bremen.de/hjaeger/pubs/ESNTutorial.pdf, GMD Report 159, 2002.

[7] L. Feldkamp, D. Prokhorov, C. Eagen, and F. Yuan, "Enhanced multi-stream Kalman filter training for recurrent neural networks," in *Nonlinear Modeling: Advanced Black-Box Techniques*, J. Suykens and J. Vandewalle, Eds. Kluwer, 1998, pp. 29–54.

[8] P.-G. Plöger, A. Arghir, T. Günther, and R. Hosseiny,

"Echo state networks for mobile robot modeling and control." in *RoboCup*, 2003, pp. 157–168.

[9] P. Berkes and L. Wiskott, "Slow feature analysis yields a rich repertoire of complex cell properties," *Cognitive Sciences EPrint Archives (CogPrints)*, vol. 2804, 2003. [Online]. Available: http://cogprints.org/2804/