Challenges in Reservoir Computing

Echo State Networks

Miguel Atencia

Universidad de Málaga

April 2024, Dauphine - PSL

1. Echo State Networks
    - Definition
    - Applications
    - Mathematical properties
2. Hyper-parameter setting
    - Empirical design
    - Automatic hyper-parameter adjustment
    - Novel proposals
3. Clustering
4. Conclusions

Time series

- Time-dependent data: $(u(t), y^{\text{target}}(t)), t = 1, \dots$
- Strong **autocorrelation**
- Series length
  - undefined (in principle)
  - variable
  - infinite (in principle)

Time series tasks

- Prediction: $y^{\text{target}}(n) = u(n+1)$
- Classification: $\vec{y}^{\text{target}} \in \{0, 1\}$
- Encoding-decoding $y^{\text{target}}(n) = F(u(n))$, e.g. *seq2seq*

Learning models

Input-output model

- Data: $(x, y^{\text{target}})_i, i = 1 \dots n$
- Model: $y(x) = \varphi(W^L \varphi(\dots(W^1 x)\dots))$
- Regression: $\min_W \|y^{\text{target}} - y(x)\|^2$
- NN are universal approximators (Cybenko, G., 1989)
- Generalization, overfitting, etc.

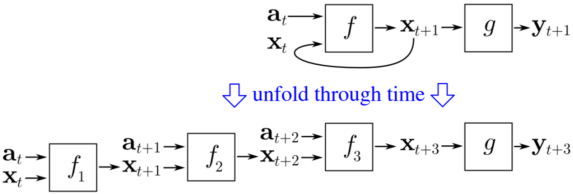
Recurrent model

- Data: $y^{\text{target}}(t), t = 1 \dots$
- Model: $x(t) = \varphi\left(W\, x(t-1) + W^{\text{fb}}\, y(t-1)\right)$
- Recurrence: $y(t) = W^o\, x(t)$

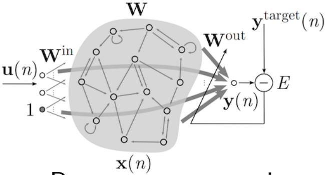
Recurrent Neural Networks

Naive extension of back-propagation learning:

- Ignore recurrence and regress at each time step
- *Unfolding* through time is equivalent to an **infinitely** Deep NN
- Computational/statistical issues
  - Computational cost
  - *Gradient vanishing*
  - *Overfitting*
- *BPTT* (1988), LSTM (1999), Transformers (2017), etc.



Echo State Networks



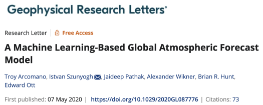
Reservoir: $x(t) \in \mathbb{R}^N$

N *arbitrary* (hyper-parameter)

```
Set constant $W$, $W^{\text{fb}}$
for $t = 1 \dots t^{\text{train}}$ do                                ▷ Regressor construction
    $u(t) = y^{\text{target}}(t)$
    $x(t+1) = \varphi\left(W\,x(t) + W^{\text{fb}}\,u(t)\right)$
end for
$X = (x(1) \dots x(t^{\text{train}}))$                                ▷ Regression
$Y = (u(1) \dots u(t^{\text{train}}))$
$\hat{W}^{\text{out}} = \operatorname{argmin}_{W^{\text{out}}} \|Y - W^{\text{out}} X\|^2$
for $t = t^{\text{train}} \dots t^{\text{pred}}$ do                  ▷ Prediction
    $x(t+1) = \varphi\left(W\,x(t) + W^{\text{fb}}\,u(t)\right)$
    $u(t+1) = \hat{W}^{\text{out}}\,x(t+1)$
end for
```

Success cases

- Weather prediction



Geophysical Research Letters

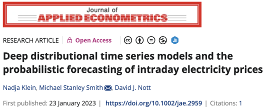
Research Letter | Free Access

A Machine Learning-Based Global Atmospheric Forecast Model

Troy Arcomano, Istvan Szunyogh, Jaideep Pathak, Alexander Wikner, Brian R. Hunt, Edward Ott

First published: 07 May 2020 | https://doi.org/10.1029/2020GL087776 | Citations: 73

- Econometrics



Journal of APPLIED ECONOMETRICS

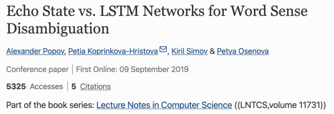
RESEARCH ARTICLE | Open Access

Deep distributional time series models and the probabilistic forecasting of intraday electricity prices

Nadja Klein, Michael Stanley Smith, David J. Nott

First published: 23 January 2023 | https://doi.org/10.1002/jae.2959 | Citations: 1

- Natural Language Processing



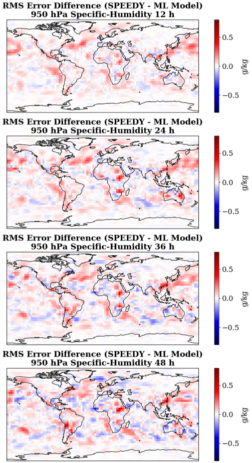
Echo State vs. LSTM Networks for Word Sense Disambiguation

Alexander Popov, Petia Koprinkova-Hristova, Kiril Simov & Petya Osenova

Conference paper | First Online: 09 September 2019

5325 Accesses | **5** Citations

Part of the book series: Lecture Notes in Computer Science ((LNTCS,volume 11731))



RMS Error Difference (SPEEDY - ML Model) 950 hPa Specific-Humidity 12 h

RMS Error Difference (SPEEDY - ML Model) 950 hPa Specific-Humidity 24 h

RMS Error Difference (SPEEDY - ML Model) 950 hPa Specific-Humidity 36 h

RMS Error Difference (SPEEDY - ML Model) 950 hPa Specific-Humidity 48 h

*ESN*s are universal approximators

Let $I_n = B(0,1) \subset \mathbb{R}^n$

- A **filter** U is a *functional* $U : (I_n)^{\mathbb{Z}} \to (\mathbb{R}^d)^{\mathbb{Z}}$
- A filter U is **causal** if $z_\tau = w_\tau$ for all $\tau \leq t$ implies $U(z)_t = U(w)_t$
- A filter U is **invariant** if it commutes with the *delay* operator: $T_\tau \circ U = U \circ T_\tau$, where $T_\tau(z)_t = z_{t-\tau}$
- An *ESN* has an associated filter U_{ESN} if its activation function φ is *squashing* (increasing, $\lim_{z\to\pm\infty} \varphi(z) = \pm 1$)

Theorem

Given a causal, invariant, continuous filter U, for all $\epsilon > 0$ there exists an *ESN* such that $\|U - U_{\text{ESN}}\|_\infty < \epsilon$.

Grigoryeva, L., Ortega, J.-P. (2018). *Echo state networks are universal*

The spectral radius of W

Recall the update equation:

$$x(t+1) = \varphi\left(W\,x(t) + W^{\text{fb}}\,u(t)\right)$$

Assume:

- No input: $u(t) = 0$
- Linear activation $\varphi(z) = z$

Then, update reduces to:

$$x(t) = W\,x(t-1)$$

Let $\rho(W)$ be the **spectral radius** de W: $\rho(W) = \max|\lambda(W)|$

$$x(t) = W^t\,x(0) = \begin{cases} \infty & \text{if } \rho(W) > 1 \\ 0 & \text{if } \rho(W) < 1 \end{cases}$$

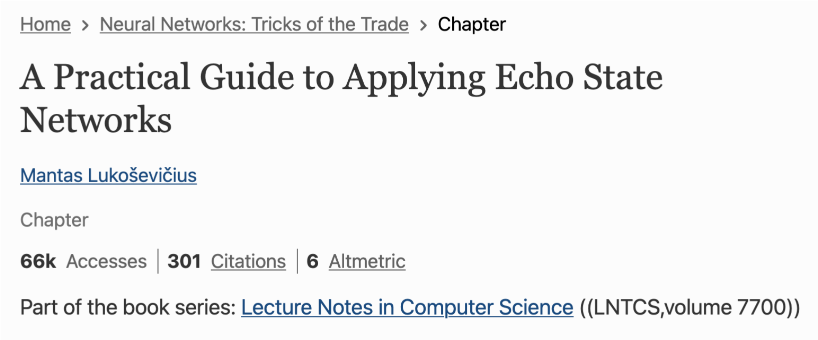
The *Echo State Property*

Definition

Uniqueness: An ESN has the *Echo State Property* (*ESP*) if $x(T) = x'(T)$, given any input sequence $u(t), t < T$ and any state sequences $x(t), x'(t), t < T$.

- Theory: Given $u(t) = 0$, an *ESN* with $\rho(W) > 1$ can not have the *ESP*
- Practice: $\rho(W) \leq 1$ is neither necessary nor sufficient for *ESP*
- *Edge of Criticality*: *intuition* of the change point from simple ($\sim$ linear) to complex ($\sim$ chaos)

Empirical ESN design



Home › Neural Networks: Tricks of the Trade › Chapter

A Practical Guide to Applying Echo State Networks

Mantas Lukoševičius

Chapter

66k Accesses | **301** Citations | **6** Altmetric

Part of the book series: Lecture Notes in Computer Science ((LNTCS,volume 7700))

Lukoševičius, M. (2012). *A Practical Guide to Applying Echo State Networks*

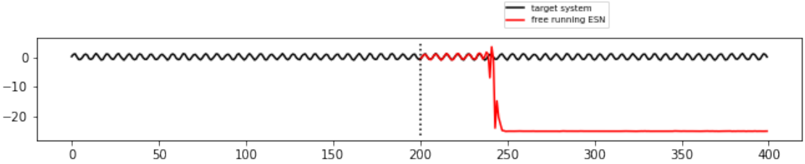
Reservoir size

Lukoševičius, M. (2012). *A Practical Guide to Applying Echo State Networks*

For challenging tasks use as big a reservoir as you can afford computationally.

N_x should be at least equal to the estimate of independent real values the reservoir has to remember from the input to solve its task.

- Data: $y^{\text{target}}(t) = \sin(t) + \epsilon\,; \quad \epsilon \sim \mathcal{U}(0, 0.3)$
- **Size**: $N = 500$



Lukoševičius, M. (2012)

> ... the bigger the reservoir, the better the obtainable performance, *provided appropriate regularization measures are taken against overfitting*

Spectral radius

Lukoševičius, M. (2012). *A Practical Guide to Applying Echo State Networks*

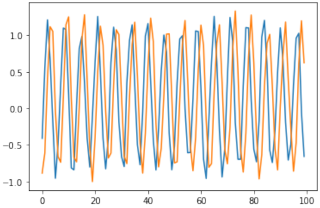
$\rho(\mathbf{W}) < 1$ ensures echo state property in most situations.

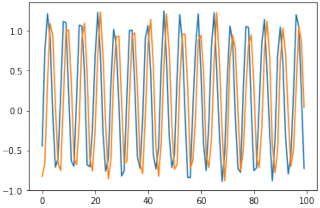
The spectral radius should be greater in tasks requiring longer memory of the input.

Note: input **scaling** required: $y^{\text{target}}(t) = 10\,(\sin(n) + \epsilon)\,; \quad \epsilon \sim \mathcal{U}(0, 0.3)$

$\rho(W) = 0.99\,; \quad E_{test} = 0.391$

$\rho(W) = 1.5\,; \quad E_{test} = 0.207$





Even $\rho(W) = 5$ has been successfully used !!

Self-normalized activation

Update equations

$$x(t) = \varphi\left(W\,x(t-1) + W^{\mathrm{fb}}\,u(t-1)\right)$$
$$y(t) = W^{\mathrm{out}}\,x(t)$$

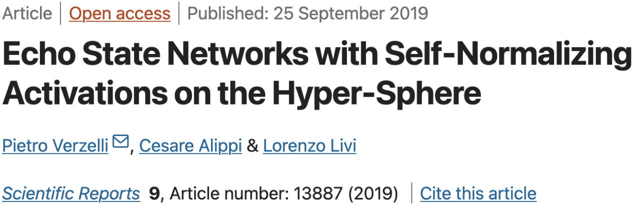
Activation function φ

Projection on hyper-sphere of radius r: $\mathbb{S}_r^{N-1} = \{p \in \mathbb{R}^N, \|p\| = r\}$

$$x(t) \leftarrow \frac{r}{\|x(t)\|}\,x(t)$$

Verzelli, P. et al. (2019). *Echo State Networks with Self-Normalizing Activations on the Hyper-Sphere*

Theoretical results



Article | Open access | Published: 25 September 2019

Echo State Networks with Self-Normalizing Activations on the Hyper-Sphere

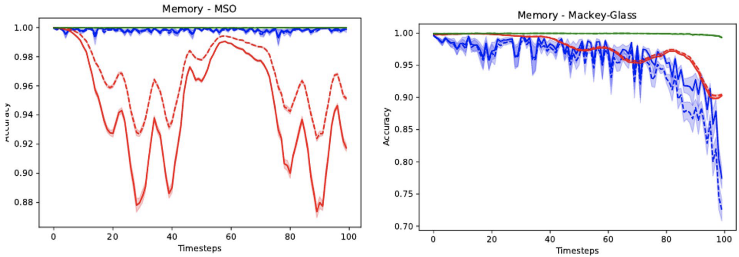
Pietro Verzelli ✉, Cesare Alippi & Lorenzo Livi

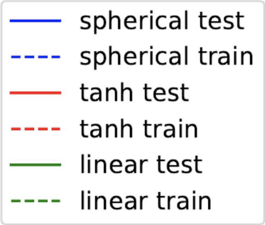
Scientific Reports **9**, Article number: 13887 (2019) | Cite this article

The *ESN* with hyper-sphere projection:

- has the *Echo State Property*
- is Universal Approximator (φ is *squashing*)
- works on the *Edge of Criticality*
    - The maximum Lyapunov exponent is 0
    - No chaotic behaviour can occur

Empirical results

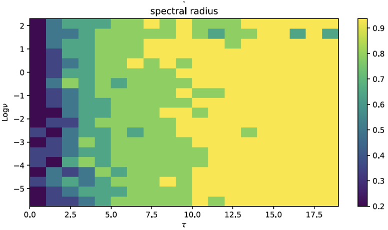




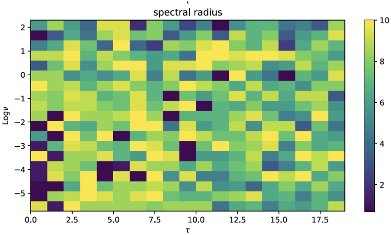
- *MSO* = sum of sinusoid functions
- $\rho(W) = 15$ in the spherical *ESN*

Memory-nonlinearity trade-off

- $y(n) = \sin(\nu\, u(k - \tau))$
- $\tau \sim$ Memory
- $\nu \sim$ Nonlinearity



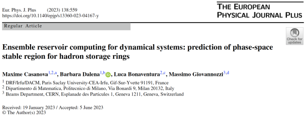
Optimal $\rho(W)$
$\varphi = \tanh$



Optimal $\rho(W)$
$\varphi =$ projection

Many other ideas

- Ensembles



Eur. Phys. J. Plus (2023) 138:559
https://doi.org/10.1140/epjp/s13360-023-04167-y

THE EUROPEAN
PHYSICAL JOURNAL PLUS

Regular Article

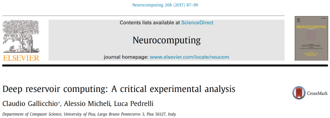
Check for updates

Ensemble reservoir computing for dynamical systems: prediction of phase-space stable region for hadron storage rings

Maxime Casanova[1,2,a], **Barbara Dalena**[1,b], **Luca Bonaventura**[2,c], **Massimo Giovannozzi**[3,d]
[1] DRF/Irfu/DACM, Paris Saclay University-CEA-Irfu, Gif-Sur-Yvette 91191, France
[2] Dipartimento di Matematica, Politecnico di Milano, Via Bonardi 9, Milan 20132, Italy
[3] Beams Department, CERN, Esplanade des Particules 1, Geneva 1211, Geneva, Switzerland

Received: 19 January 2023 / Accepted: 5 June 2023
© The Author(s) 2023

- Deep ESN, sparsity



Neurocomputing 268 (2017) 87–99

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

ELSEVIER

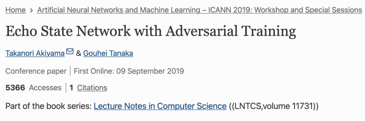
Deep reservoir computing: A critical experimental analysis

CrossMark

Claudio Gallicchio*, Alessio Micheli, Luca Pedrelli

Department of Computer Science, University of Pisa, Largo Bruno Pontecorvo 3, Pisa 56127, Italy

- Learning models



Home › Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions

Echo State Network with Adversarial Training

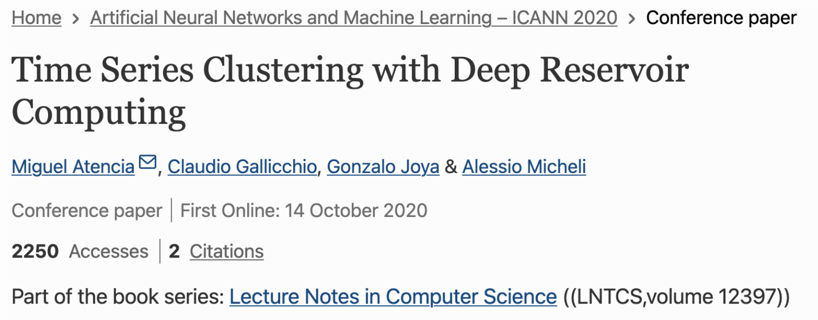
Takanori Akiyama & Gouhei Tanaka

Conference paper | First Online: 09 September 2019

5366 Accesses | **1** Citations

Part of the book series: Lecture Notes in Computer Science ((LNTCS,volume 11731))

Clustering



Home › Artificial Neural Networks and Machine Learning – ICANN 2020 › Conference paper

Time Series Clustering with Deep Reservoir Computing

Miguel Atencia, Claudio Gallicchio, Gonzalo Joya & Alessio Micheli

Conference paper | First Online: 14 October 2020

2250 Accesses | **2** Citations

Part of the book series: Lecture Notes in Computer Science ((LNTCS,volume 12397))

Ignore learning and pay attention only to dynamics:

$$u(t) = y^{\text{target}}(t)$$

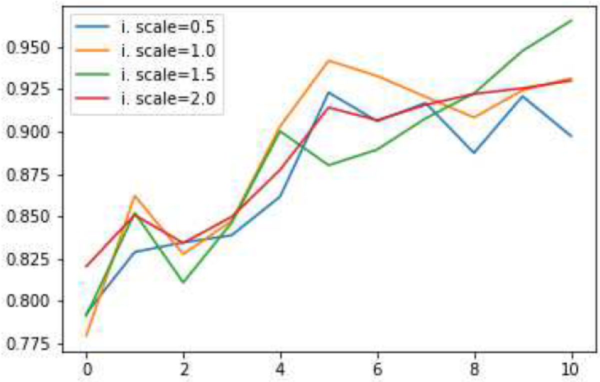
$$x(t+1) = \varphi\left(W\,x(t) + W^{\text{fb}}\,u(t)\right)$$

Dynamic clustering

Algorithm 1 Dynamic clustering through evolution of the RC model.

```
Require: Dataset of $n$ time series $\mathbf{u}_j$ with lengths $l_j$, $j = 1 \dots n$.
Ensure: $k$ centroids
Initialise weight matrices $\mathbf{U}^{(i)}$, $\mathbf{W}^{(i)}$, $i = 1 \dots L$, and replicate $n$ identical instances
Initialise all instances states $\mathbf{h}^{(i)}(0) = 0$, $i = 1 \dots L$
for $t = 1$ to $\max_j l_j$ do
    for $j = 1$ to $n$ do
        if $t \leq l_j$ then
            Update the corresponding ESN instance by Equation (1)
        end if
    end for
    if $t = 1$ then
        Initialize centroids
    else
        Set initial centroids to centroids resulting from step $t - 1$
    end if
    Build the dataset $\mathbf{Y}(t)$ of $n$ reservoir states, where $\mathbf{Y}_{i\cdot}(t) = \left(\mathbf{h}^{(1)}, \dots \mathbf{h}^{(L)}\right)$
    Compute centroids at step $t$ from clustering of dataset $\mathbf{Y}(t)$
end for
```

Results on Deep ESN Dynamic Clustering



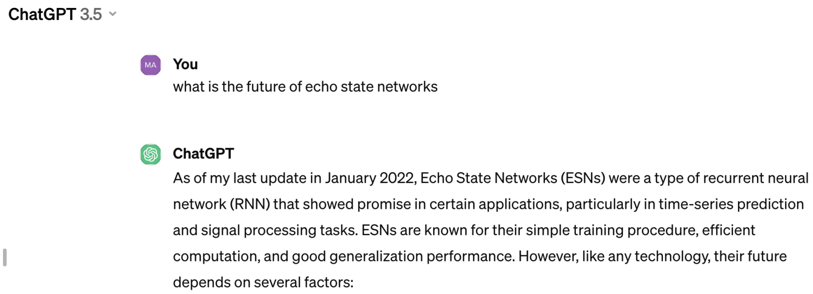
Silhouette coefficient of clustering by layer
(for several values of input scaling).

Q & A

Personal conclusions:

Answer	**Question**	Answer
Successful applications	Do ESN work?	Sort of
Theoretical results	Why do ESN work?	No idea
Easy implementation Low computational cost	How do ESN work?	Frustrating hyperparameter setting

The future of *ESN*



You

what is the future of echo state networks

ChatGPT

As of my last update in January 2022, Echo State Networks (ESNs) were a type of recurrent neural network (RNN) that showed promise in certain applications, particularly in time-series prediction and signal processing tasks. ESNs are known for their simple training procedure, efficient computation, and good generalization performance. However, like any technology, their future depends on several factors:

- Application Domains
- Theoretical Understanding
- Competition from Other Models

Thanks to:

- CEREMADE (Madalina, Katia, and everyone)
- Dauphine-PSL

Miguel Atencia
`matencia@uma.es`