



ANTISYMMETRICRNN: A DYNAMICAL SYSTEM VIEW ON RECURRENT NEURAL NETWORKS


<https://arxiv.org/abs/1902.09689>

Luca Palumbo

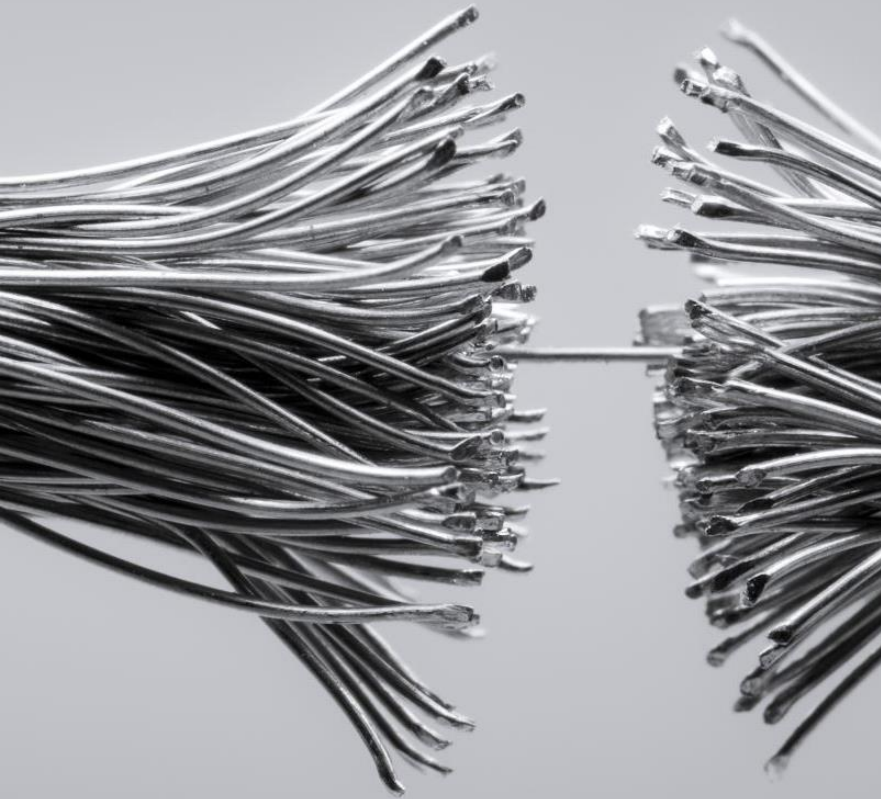
639750

INTRODUCTION

Recurrent Neural Networks (RNNs) suffer from the vanishing gradient problem, which interfere with their ability to learn long-term dependencies in sequential data. Several solutions have been proposed — such as gated architectures or constraining the weight matrix to be orthogonal — but these approaches only partially address the issue.



This paper proposes a novel perspective based on **dynamical systems theory**: by designing the RNN as a stable differential equation, the model naturally preserves information over time, improving its ability to capture long-term dependencies.



RNN-ODE CONNECTION

- Considers the following ODE: $h'(t) = \tanh(Wh(t))$
- Now we can discretize it by means of Forward Euler Method: $h_t = h_{t-1} + \epsilon \tanh(Wh_{t-1})$
- This equation resemble a recurrent neural network without input data. (W are the model parameters, ϵ is an hyperparameter, h_t is the hidden state)

MATHEMATICAL PILLS

- Stability Definition: *a solution $h(t)$ of an ODE with initial condition $h(0)$ is stable if $\forall \epsilon > 0 \exists \delta > 0$ such that any other solution $\tilde{h}(t)$ of the ODE with initial condition $\tilde{h}(0)$ satisfying $|h(0) - \tilde{h}(0)| < \delta$ also satisfies $|h(t) - \tilde{h}(t)| < \epsilon$*
 - Proposition: *The solution of an ODE is stable if $\max_i \operatorname{Re}(\lambda_i(J(t))) \leq 0$, where $\operatorname{Re}()$ denotes the real part of a complex number, $\lambda_i(J(t))$ denotes the eigenvalues of the jacobian of f in the equation $h'(t) = f(h(t))$*
 - Stability alone does not suffice to capture long-term dependencies: if $\operatorname{Re}(\lambda_i(J(t))) \ll 0 \forall i$ the correspondent RNN will lead to catastrophic forgetting.
 - Under the **Critical Criterion** $\operatorname{Re}(\lambda_i(J(t))) \approx 0$ the system preserves long-term dependencies of the inputs while being stable.
 - Antisymmetrical Matrix definition and property: *a matrix M is antisymmetric if $M^T = -M$. Eigenvalues of antisymmetric matrix are all imaginary: $\operatorname{Re}(\lambda_i(M)) = 0$*
-

FULL MODEL DESCRIPTION

$$z_t = \sigma \left((W_h - W_h^T - \gamma 1)h_{t-1} + V_z x_t + b_z \right)$$

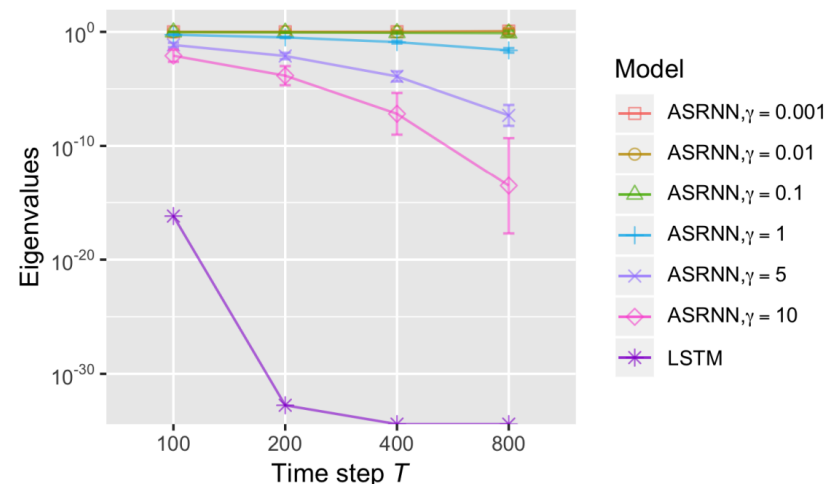
$$h_t = h_{t-1} + \epsilon z_t * \tanh \left((W_h - W_h^T - \gamma 1)h_{t-1} + V_h x_t + b_h \right)$$

- $W_h - W_h^T$ is the antisymmetric weight matrix that make the critical criterion satisfied
- $-\gamma 1$ is required to make the discretized ode stable – a stable ode doesn't necessary imply a stable discretized version
- z_t is the input gate and $*$ is the hadamard product

EXPERIMENT: NOISE PADDED CIFAR-10

- Authors inputted each row of a CIFAR-10 image at every time step. After the first 32 time steps, they inputted independent standard Gaussian noise for the remaining time steps. The total number of time steps is set to $T = 1000$. In other words, only the **first 32 time steps of input contain salient information, all remaining 968 time steps are merely random noise**. For a model to correctly classify an input image, **it has to remember the information from a long time ago**.
- An additional set of experiments varying the length of noise padding was performed. In the graph we can visualize the mean and standard deviation of the eigenvalues of the end-to-end Jacobian matrices. Unitary eigenvalues, i.e., mean close to 1 and standard deviation close to 0, indicate non-exploding and non-vanishing gradients

method	pixel-by-pixel	noise padded	# units	# params
LSTM	59.7%	11.6%	128	69k
Ablation model	54.6%	46.2%	196	42k
AntisymmetricRNN	58.7%	48.3%	256	36k
AntisymmetricRNN w/ gating	62.2%	54.7%	256	37k



CONCLUSION



This research drew connections between RNNs and the ordinary differential equation theory



The model AntisymmetricRNN was proposed which is a discretization of ODEs that satisfy the critical criterion with competitive performance



Hopefully this work will inspire future research in both communities of RNN and ODE.

THANKS FOR
YOUR
ATTENTION

