Gated Recurrent Neural Networks with Weighted Time-Delay Feedback

N. Benjamin Erichson (ICSI & LBNL), Soon Hoe Lim (KTH & Nordita), and Michael Mahoney (UC Berkeley, ICSI & LBNL)

SCAN ME

Introduction

- **Design:** We introduce a novel gated recurrent unit, the τ-GRU, which incorporates a weighted time-delay feedback mechanism to mitigate the vanishing gradient problem. This architecture is derived by numerically discretizing a carefully designed continuous-time delay differential equation (DDE).
- **Theory:** We show that the continuous-time τ-GRU model admits a unique solution. Furthermore, we provide both intuition and theoretical analysis to demonstrate how the delay term in τ-GRU can act as a memory buffer, helping to alleviate the vanishing gradient problem and enhancing the model's ability to retain long-range dependencies.
- **Experiments:** We show that τ-GRU converges faster during training and achieves improved generalization performance, outperforming existing *nonlinear* RNN models across a diverse set of challenging tasks.

Delay Differential Equations (DDEs)

DDEs are a class of dynamical systems in which a feedback term is introduced to adjust the system non-instantaneously, resulting in delays in time:

$$\dot{h} = F(t, h(t), h(t-\tau)), \tag{1}$$

with $\tau > 0$, where F is a continuous function.

- Need to *specify an initial function* to describe the behavior of the system prior to the initial time 0: it would be a function ϕ defined on $[-\tau, 0]$
- More precisely, the DDE is a functional differential equation with the state space $C := C([-\tau, 0], \mathbb{R}^d)$. This state space is the Banach space of continuous functions from $[-\tau, 0]$ into $\mathbb{R}^d$, with the topology of uniform convergence. It is equipped with the norm $\|\phi\| := \sup\{|\phi(\theta)| : \theta \in [-\tau, 0]\}$
- In contrast to the ODEs (with $\tau = 0$) whose state space is finite-dimensional, DDEs are generally infinite-dimensional dynamical systems

From DDEs to Continuous-Time RNNs

- Use input-driven nonlinear DDEs to model the dynamics of the hidden states:

$$\frac{dh(t)}{dt} = f(h(t), h(t-\tau), x(t); \theta),$$

where τ is a constant that indicates the delay (time-lag). Here, the time derivative is described by a function $f : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ that explicitly depends on states from the past

- The basic form of a time-delayed RNN is

$$\dot{h} = \sigma(W_1 h(t) + W_2 h(t-\tau) + Ux(t)) - h(t), \tag{2}$$

for $t \geq 0$, and $h(t) = 0$ for $t \in [-\tau, 0]$, with the output $y(t) = Vh(t)$. In this expression, $h \in \mathbb{R}^d$ denotes the hidden states, $f : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ is a nonlinear function, and $\sigma : \mathbb{R} \to (-1, 1)$ denotes the tanh activation function applied component-wise. The matrices $W_1, W_2 \in \mathbb{R}^{d\times d}$, $U \in \mathbb{R}^{d\times p}$ and $V \in \mathbb{R}^{q\times d}$ are learnable parameters, and $\tau \geq 0$ denotes the discrete time-lag

- To better represent a large number of scales, consider a time warping function $c : \mathbb{R}^d \to \mathbb{R}^d$ which we define to be a parametric function $c(t)$, and denoting $t_\tau := t - \tau$:

$$\dot{h} = \frac{dc(t)}{dt}\left[\sigma(W_1 h(t) + W_2 h(t_\tau) + U_1 x(t)) - h(t)\right] \tag{3}$$

- Now, we need a learnable function to model $\frac{dc(t)}{dt}$. A natural choice is to consider a standard gating function, which is a universal approximator, taking the form

$$\frac{dc(t)}{dt} = \hat{\sigma}(W_3 h_t + U_3 x_t) =: g(t), \tag{4}$$

where $W_3 \in \mathbb{R}^{d\times d}$ and $U_3 \in \mathbb{R}^{d\times p}$ are learnable parameters, and where $\hat{\sigma} : \mathbb{R} \to (0, 1)$ is the component-wise sigmoid function

From Continuous-Time to Discrete-Time RNNs

- Using the explicit forward Euler scheme and choosing $\Delta t = 1$ gives:

$$h_{n+1} = (1 - g_n) \odot h_n + g_n \odot \sigma(W_1 h_n + W_2 h_l + U h_n) \tag{7}$$

Use a mixture of a standard recurrent unit and a delay recurrent unit: replace the σ in Eq. (7) by

$$u_n + a_n \odot z_n, \tag{8}$$

so that we yield a new GRU that takes the form

$$h_{n+1} = (1 - g_n) \odot h_n + g_n \odot (u_n + a_n \odot z_n) \tag{9}$$

c.f. sigmoidal coupling used in Hodgkin-Huxley type neural models

τ-GRU

Continuous-time formulation of τ-GRU:

$$\frac{d\,\mathbf{h}(t)}{dt} = \underbrace{g(\mathbf{h}(t), \mathbf{x}(t))}_{\text{gating}}\left(\underbrace{u(\mathbf{h}(t), \mathbf{x}(t))}_{\text{instantaneous dynamics}} + \underbrace{a(\mathbf{h}(t), \mathbf{x}(t)) \odot z(\mathbf{h}(t-\tau), \mathbf{x}(t))}_{\text{weighted time-delayed feedback}} - \mathbf{h}(t)\right) \tag{1}$$

Discrete-time formulation of τ-GRU:

$$\mathbf{h}_{n+1} = (1 - \mathbf{g}_n) \odot \mathbf{h}_n + \mathbf{g}_n \odot (\mathbf{u}_n + \mathbf{a}_n \odot \mathbf{z}_n) \tag{2}$$

with

$$\mathbf{u}_n = u(\mathbf{h}_n, \mathbf{x}_n) := \tanh(\mathbf{W}_1\mathbf{h}_n + \mathbf{U}_1\mathbf{x}_n) \tag{3}$$
$$\mathbf{z}_n = z(\mathbf{h}_l, \mathbf{x}_n) := \tanh(\mathbf{W}_2\mathbf{h}_l + \mathbf{U}_2\mathbf{x}_n) \tag{4}$$
$$\mathbf{g}_n = g(\mathbf{h}_n, \mathbf{x}_n) := \mathrm{sigmoid}(\mathbf{W}_3\mathbf{h}_n + \mathbf{U}_3\mathbf{x}_n) \tag{5}$$
$$\mathbf{a}_n = a(\mathbf{h}_n, \mathbf{x}_n) := \mathrm{sigmoid}(\mathbf{W}_4\mathbf{h}_n + \mathbf{U}_4\mathbf{x}_n) \tag{6}$$

input	$\mathbf{x}$	$\mathbb{R}^p$
time index	t	$\mathbb{R}$
time delay	τ	$\mathbb{R}$
hidden state	$\mathbf{h}$	$\mathbb{R}^d$
hidden-to-hidden matrix	$\mathbf{W}_i$	$\mathbb{R}^{d\times d}$
input-to-hidden matrix	$\mathbf{U}_i$	$\mathbb{R}^{d\times p}$
decoder matrix	$\mathbf{V}$	$\mathbb{R}^{q\times d}$

$$\mathbf{h}_n \approx \mathbf{h}(t_n),\ t_n = n\Delta t,\ n = 0, 1, \ldots$$
$$l := n - \lfloor \tau/\Delta t \rfloor$$

Result 1: τ-GRU Has a Unique Solution

Theorem (Existence and uniqueness of solution for continuous-time τ-GRU)

Let $t_0 \in \mathbb{R}$ and $\phi \in C$ be given. There exists a unique solution $h(t) = h(t, \phi)$ of Eq. (1), defined on $[t_0 - \tau, t_0 + A]$ for any $A > 0$. In particular, the solution exists for all $t \geq t_0$, and

$$\|h_t(\phi) - h_t(\psi)\| \leq \|\phi - \psi\| e^{K(t-t_0)}, \tag{13}$$

for all $t \geq t_0$, where $K = 1 + \|W_1\| + \|W_2\| + \|W_4\|/4$.

Result 2: The Delay Term Mitigates Vanishing Gradient Problem

Proposition

Consider the linear time-delayed RNN whose hidden states are described by the update equation:

$$h_{n+1} = Ah_n + Bh_{n-m} + Cu_n, \ \ n = 0, 1, \ldots, \tag{17}$$

and $h_n = 0$ for $n = -m, -m+1, \ldots, 0$ with $m > 0$.
Then, assuming that A and B commute, we have:

$$\frac{\partial h_{n+1}}{\partial u_i} = A^{n-i}C, \tag{18}$$

for $n = 0, 1, \ldots, m$, $i = 0, \ldots, n$, and

dampen the exponential decay!

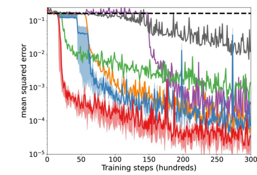
$$\frac{\partial h_{m+1+j}}{\partial u_i} = A^{m+j-i}C + \delta_{i,j-1}BC + 2\delta_{i,j-2}ABC + 3\delta_{i,j-3}A^2BC + \cdots + j\delta_{i,0}A^{j-1}BC, \tag{19}$$

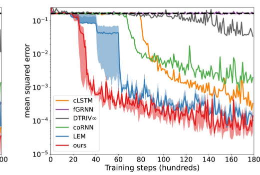
for $j = 1, 2, \ldots, m+1$, $i = 0, 1, \ldots, m+j$, where $\delta_{i,j}$ denotes the Kronecker delta.

*Similar gradient bounds can also be derived for the fully nonlinear RNN model under suitable assumptions on the activation functions.

Empirical Results

The adding task:





Sequential image classification:

Model	sMNIST	psMNIST	# units	# parms	sCIFAR	nCIFAR	# units	# parms
LSTM (Kag and Saligrama, 2021)	97.8	92.6	128	68k	59.7	11.6	128	69k / 117k
r-LSTM (Trinh et al., 2018)	98.4	95.2	-	100K	72.2	-	-	101k / -
chrono-LSTM (Rusch et al., 2022)	98.9	94.6	128	68k	-	55.9	128	- / 116k
Antisym. RNN (Chang et al., 2018)	98.0	95.8	128	10k	62.2	54.7	256	37k / 37k
Lipschitz RNN (Erichson et al., 2020)	99.4	96.3	128	34k	64.2	59.0	256	134k / 158k
expRNN (L-Casado and M-Rubio, 2019)	98.4	96.2	360	68k	-	49.0	128	- / 47k
iRNN (Kag et al., 2020)	98.1	95.6	128	8k	-	54.5	128	- / 12k
TARNN (Kag and Saligrama, 2021)	98.9	97.1	128	68k	-	59.1	128	- / 100K
Dilated GRU (Chang et al., 2017)	99.2	94.6	-	130k	-	-	-	- / -
coRNN (Rusch and Mishra, 2021)	99.3	96.6	128	34k	-	59.0	128	- / 46k
LEM (Rusch et al., 2022)	**99.5**	96.6	128	68k	-	60.5	128	- / 117k
Delay GRU (Eq. (12))	98.7	94.1	128	51k	56.1	53.7	128	52k / 75k
τ-GRU (ours)	99.4	97.3	128	68k	**74.9**	**62.7**	128	69k / 117k

Simple delayed RNN: $\mathbf{h}_{n+1} = (1-\mathbf{g}_n) \odot \mathbf{h}_n + \mathbf{g}_n \odot \sigma(\mathbf{W}_1\mathbf{h}_n + \mathbf{W}_2\mathbf{h}_l + \mathbf{U}\mathbf{x}_n). \quad (12)$

With ablation parameters: $\mathbf{h}_{t+1} = (1 - \mathbf{g}_t) \odot \mathbf{h}_t + \mathbf{g}_t \odot (\beta \cdot \mathbf{u}_t + \alpha \cdot \mathbf{a}_t \odot \mathbf{z}_t)$

Learning climate dynamics (ENSO):

Model	MSE ($\times 10^{-2}$)	# units	# parameter
Vanilla RNN	0.45	16	0.3k
LSTM	0.92	16	1.2k
GRU	0.53	16	0.9k
Lipschitz RNN	10.6	16	0.6k
coRNN	4.00	16	0.6k
LEM	0.31	16	1.2k
ablation ($\alpha = 0$)	0.31	16	0.6k
ablation ($\beta = 0$)	0.38	16	0.9k
τ-GRU (ours)	**0.17**	16	1.2k

Frequency classification:

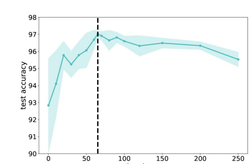
Model	No noise	With noise
Tanh-RNN	97.1%	35.6%
LSTM	100.0%	39.4%
LSTM (w/o forget gate)	99.0%	19.4%
LEM	96.0%	54.1%
SSM-S4D (1 layer)	67.5%	66.4%
SSM-S4D (4 layers)	68.9%	67.2%
GRU (no delay, ablation)	95.0%	57.7%
GRU (with delay, ours)	**100.0%**	**99.1%**

Human activity recognition (HAR2):

Model	Test Acc. (%)	# units	# param
GRU (Kusupati et al., 2018)	93.6	75	19k
LSTM (Kag et al., 2020)	93.7	64	19k
FastRNN (Kusupati et al., 2018)	94.5	80	7k
FastGRNN (Kusupati et al., 2018)	95.6	80	7k
AsymRNN (Kag et al., 2020)	93.2	120	8k
iRNN (Kag et al., 2020)	96.4	64	4k
DIRNN (Zhang et al., 2021)	96.5	64	-
coRNN (Rusch and Mishra, 2021)	97.2	64	9k
LipschitzRNN	95.4	64	9k
LEM	97.1	64	19k
τ-GRU (ours)	**97.4**	64	19k

Ablation study on psMNIST:

Model	α	β	τ	$\mathbf{a}_t$	Accuracy (%)
ablation	0	1	-	yes	94.6
ablation	1	0	65	yes	94.9
ablation	1	1	0	yes	95.1
ablation	1	1	20	yes	96.4
ablation	1	1	65	no	96.8
τ-GRU (ours)	1	1	65	yes	**97.3**



Sentiment analysis (IMDB):

Model	Test Acc. (%)	# units	# param
LSTM (Campos et al., 2018)	86.8	128	220k
Skip LSTM (Campos et al., 2018)	86.6	128	220k
GRU (Campos et al., 2018)	86.2	128	164k
Skip GRU (Campos et al., 2018)	86.6	128	164k
ReLU GRU (Dey and Salem, 2017)	84.8	128	99k
coRNN (Rusch and Mishra, 2021)	87.4	128	46k
LEM	88.1	128	220k
τ-GRU (ours)	**88.7**	128	220k