RUB

RUHR-UNIVERSITÄT BOCHUM

ACTIVITY-SPARSE INFERENCE AND LEARNING IN RECURRENT NEURAL NETWORKS

Anand Subramoney

INI

Looking beyond biology based models for neuromorphic computing

- Spiking neural networks (SNNs) were developed as models of biological neurons
- SNNs have become the canonical model for neuromorphic computing
- BUT, focus of neuromorphic devices is shifting further towards deep learning applications with higher expectations of task performance
- Are these biologically inspired spiking neural networks optimal for neuromorphic computing?
- **We need to design deep learning architectures *ab initio* for neuromorphic computing**
  - By distilling the essential advantageous properties of these biological models

What are the key properties of SNNs?

aka desiderata for neuromorphic architectures

- **Sparsity –** can be in both time (activity) and space (parameters)
  - Activity sparsity – activity transmitted only when needed
  - Parameter sparsity – activity transmitted only to units that need them

- **Event based communication –** communication happens only through discrete events between units
  - Combined with activity sparsity, units only need to update state on incoming event

- Asynchrony – no shared clock signal

- Other properties?

arXiv.2206.06178

Event-based Gated Recurrent Unit (EGRU)

Based on GRU, a very performant recurrent architecture for deep learning.

GRU Equations:

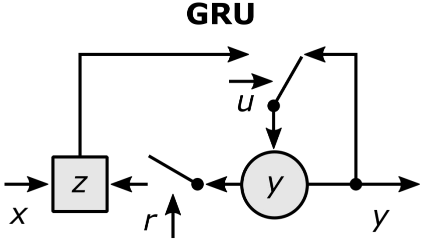
Update gate

Reset gate

$$\mathbf{u}^{\langle t \rangle} = \sigma\left(\mathbf{W}_u\left[\mathbf{x}^{\langle t \rangle}, \mathbf{y}^{\langle t-1 \rangle}\right] + \mathbf{b}_u\right), \quad \mathbf{r}^{\langle t \rangle} = \sigma\left(\mathbf{W}_r\left[\mathbf{x}^{\langle t \rangle}, \mathbf{y}^{\langle t-1 \rangle}\right] + \mathbf{b}_r\right),$$

$$\mathbf{z}^{\langle t \rangle} = g\left(\mathbf{W}_z\left[\mathbf{x}^{\langle t \rangle}, \mathbf{r}^{\langle t \rangle} \odot \mathbf{y}^{\langle t-1 \rangle}\right] + \mathbf{b}_z\right), \quad \boxed{\mathbf{y}^{\langle t \rangle} = \mathbf{u}^{\langle t \rangle} \odot \mathbf{z}^{\langle t \rangle} + (1 - \mathbf{u}^{\langle t \rangle}) \odot \mathbf{y}^{\langle t-1 \rangle},}$$

Output



Add event-generating mechanism:

Output

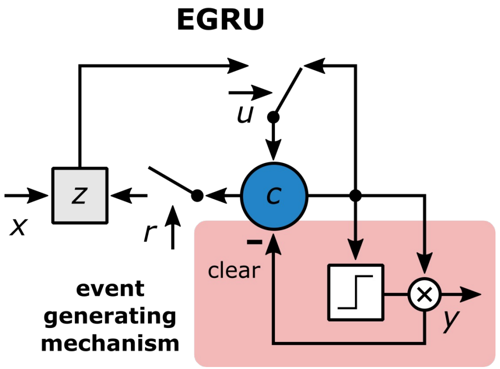
Cell state

$$\boxed{y_i^{\langle t \rangle} = c_i^{\langle t \rangle} H\left(c_i^{\langle t \rangle} - \vartheta_i\right)} \quad \text{with} \quad c_i^{\langle t \rangle} = u_i^{\langle t \rangle} z_i^{\langle t \rangle} + (1 - u_i^{\langle t \rangle}) c_i^{\langle t-1 \rangle} - y_i^{\langle t-1 \rangle},$$

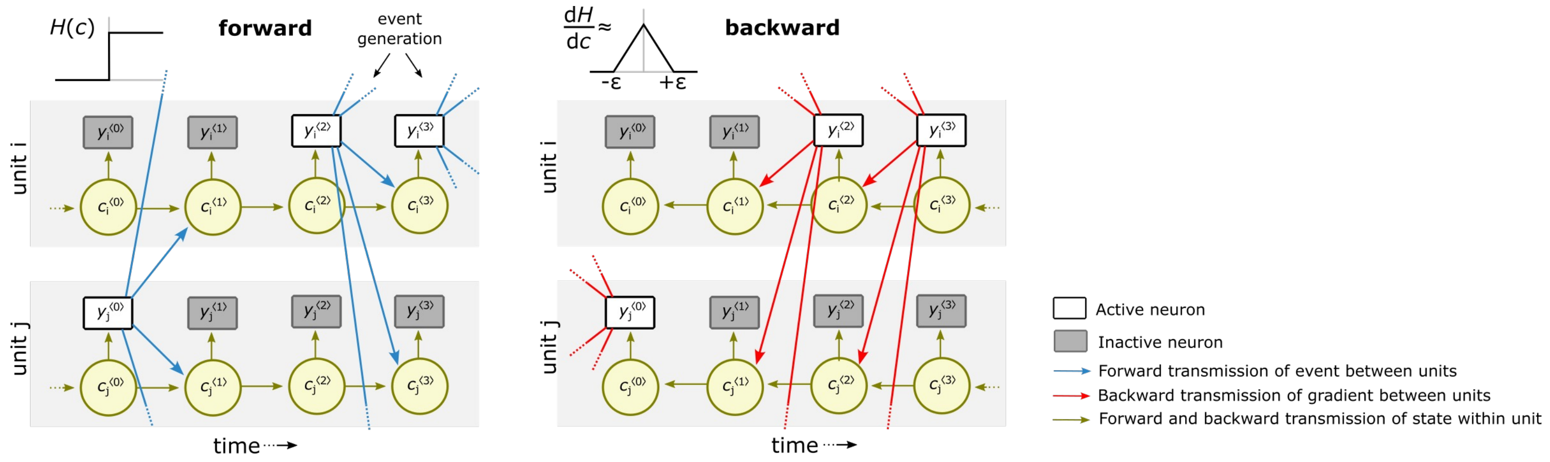
Heaviside step function

Threshold

Reset



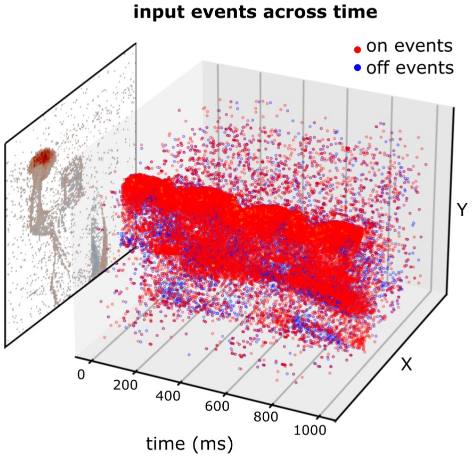
Learning in EGRU

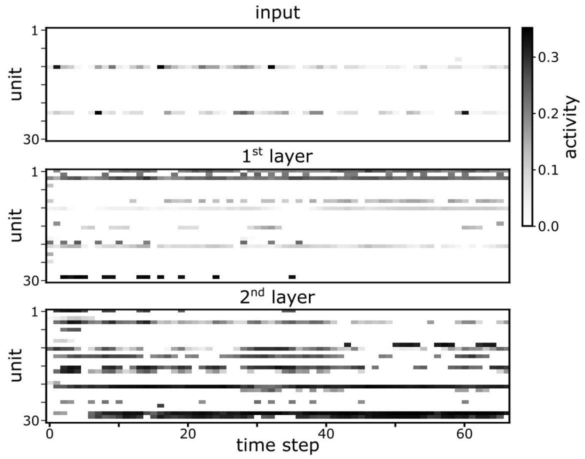


- Use a pseudo-derivative for the non-differentiable threshold function
- Choosing appropriate pseudo-derivative makes BPTT backward pass sparse
- Beyond the support of the pseudo-derivative, gradients are not backpropagated.

Parameter updates from backpropagation-through-time (BPTT) also sparse!

Results: DVS 128 gesture recognition

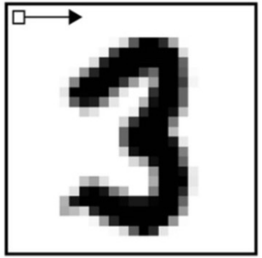




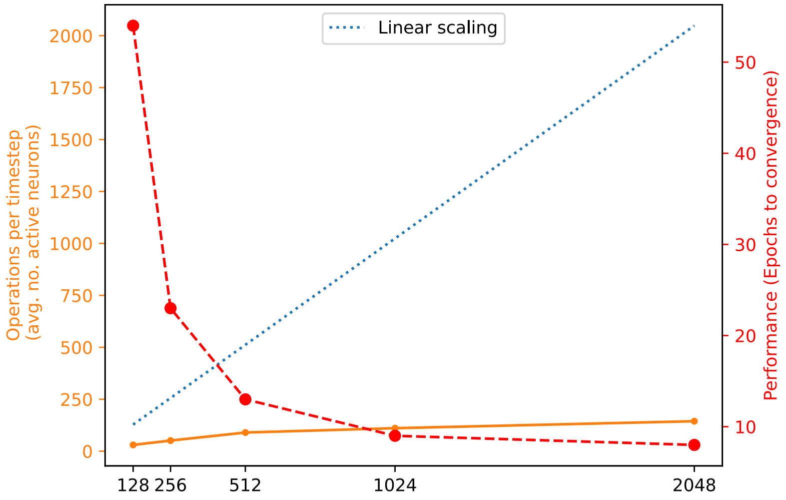
reference	architecture (# units)	para-meters	effective MAC	accu-racy	activity sparsity	backward sparsity
He et al. [23]	LSTM (512)	7.35M	7.34M	86.81%	-	-
Innocenti et al. [31]	AlexNet+LSTM+DA	9.99M	638.25M	97.73%	-	-
ours	GRU (1024)	15.75M	15.73M	88.07%	0%	-
ours	**EGRU** (512)	5.51M	4.19M	88.02%	83.79%	53.55%
ours	**EGRU** (1024)	15.75M	10.54M	90.22%	82.53%	56.63%
ours	**EGRU**+DA (1024)	15.75M	10.77M	97.13%	78.77%	58.20%

RUHR UNIVERSITÄT BOCHUM

Results: sequential MNIST image classification



reference	architecture (# units)	parameters	effective MAC	test accuracy	activity sparsity
Rusch and Mishra [55]	coRNN (256)	134K	262K	99.4%	-
Gu et al. [22]	LSTM (512)	1M	1M	98.8%	-
ours	GRU (590)	1M	1M	98.8%	-
ours	**EGRU** (590)	1M	226K	98.3%	72.1%

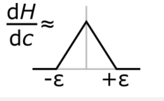


Scaling:

- Operations per timestep remains almost constant with network size for given task
- Larger networks converge faster

RUHR UNIVERSITÄT BOCHUM

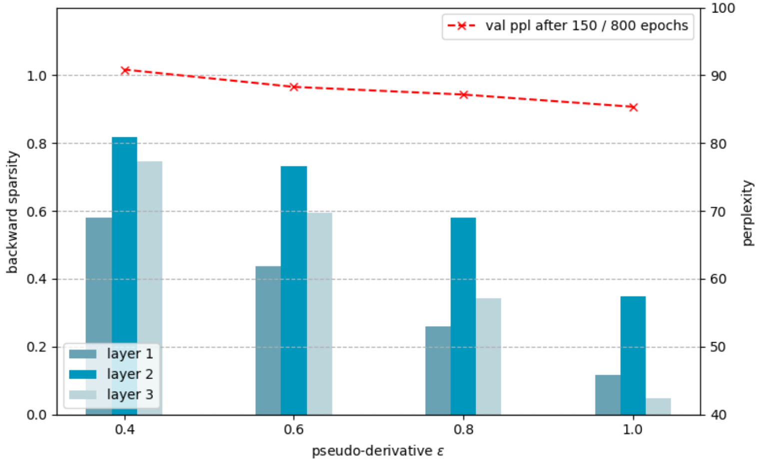
Results: Language Modelling



Dataset: PennTreeBank

Metric: Perplexity (lower is better)

reference	architecture (# units)	para-meters	effective MAC	validation	test	activity sparsity
Gal et al. [18]	Variational LSTM	24M	-	77.3	75.0	-
Melis et al. [45]	1 layer LSTM	24M	-	61.8	59.6	-
Merity et al. [46]	AWD-LSTM	24M	24M	60.0	57.3	-
ours	GRU (1350)	24M	24M	75.1	71.1	-
ours	**EGRU** (1350)	24M	5.4M	68.6	65.6	87.3%
ours	**EGRU** (2700)	76M	8.4M	69.7	66.6	91.2%



RUHR UNIVERSITÄT BOCHUM RUB

EGRU can also be written in continuous time

GRU equations are forward Euler equations of a continuous time model

$$
\begin{aligned}
& \mathbf{y}^{\langle t \rangle} = \mathbf{u}^{\langle t \rangle} \odot \mathbf{z}^{\langle t \rangle} + (1 - \mathbf{u}^{\langle t \rangle}) \odot \mathbf{y}^{\langle t-1 \rangle} \\
\Longleftrightarrow \quad & \mathbf{y}^{\langle t \rangle} - \mathbf{y}^{\langle t-1 \rangle} = -\mathbf{u}^{\langle t \rangle} \odot \mathbf{y}^{\langle t-1 \rangle} + \mathbf{u}^{\langle t \rangle} \odot \mathbf{z}^{\langle t \rangle} \\
\xrightarrow{\text{limit}} \quad & \dot{\mathbf{y}}(t) = -\mathbf{u}(t) \odot (\mathbf{y}(t) - \mathbf{z}(t))
\end{aligned}
$$

Adding activations ("synaptic currents") $\boldsymbol{a}$ and separating internal state $\boldsymbol{c}$ from output $\boldsymbol{y}$, we get:

$$\tau_m \dot{\mathbf{c}}(t) = \mathbf{u}(t) \odot (\mathbf{z}(t) - \mathbf{c}(t)) = F(t, \mathbf{a}_u, \mathbf{a}_r, \mathbf{a}_z, \mathbf{c}),$$

With the gates defined as:

$$\mathbf{u}(t) = \sigma(\mathbf{a}_u(t)), \quad \mathbf{r}(t) = \sigma(\mathbf{a}_r(t)), \quad \mathbf{z}(t) = g(\mathbf{a}_z(t)),$$

$$\text{with dynamics} \quad \tau_s \dot{\mathbf{a}}_{\mathrm{X}} = -\mathbf{a}_{\mathrm{X}} - \mathbf{b}_{\mathrm{X}}, \quad \mathrm{X} \in \{u, r, z\}$$

Events in continuous time EGRU

Internal event no. k is triggered when $c_{n_k}(s_k)$ reaches threshold ϑ_n at time s_k ($.^-$ and $.^+$ denote *before* and *after* event)

$$c_{n_k}^-(s_k) = \vartheta_{n_k}, \quad c_{n_k}^+(s_k) = 0$$
$$c_m^+(s_k) = c_m^-(s_k)$$

Activations ("synaptic currents") are updated as:

$$a_{\mathsf{x},m}^+(s_k) = a_{\mathsf{x},m}^-(s_k) + v_{\mathsf{x},mn_k} \times r_{n_k} \times c_{n_k}^-(s_k)$$

for $\mathsf{x} \in \{u, r, z\}$.

Input events have comparable updates

RUHR UNIVERSITÄT BOCHUM

Event-based gradient descent rule

similar to event-prop

Writing the loss as: $\mathcal{L} = \int_0^T \left[\underbrace{\ell_c(\mathbf{c}(t), t)}_{\text{loss}} + \underbrace{\boldsymbol{\lambda}_c \cdot (\tau_m \dot{\mathbf{c}}(t) - F(t, \mathbf{a}_u, \mathbf{a}_r, \mathbf{a}_z, \mathbf{c}))}_{\text{cell state adjoint}} + \sum_{\mathrm{x} \in \{u,r,z\}} \underbrace{\boldsymbol{\lambda}_{a\mathrm{x}} \cdot (\tau_s \dot{\mathbf{a}}_\mathrm{x} + \mathbf{a}_\mathrm{x} + \mathbf{b}_\mathrm{x})}_{\text{activation adjoints}} \right] dt\ .$

$\mathrm{x} \in \{u, r, z\}$

The adjoint dynamics are ODEs:

$$\tau_m \dot{\boldsymbol{\lambda}}_c = \left(\frac{\partial F}{\partial \mathbf{c}}\right)^T \boldsymbol{\lambda}_c \qquad \text{with} \quad \boldsymbol{\lambda}_c(T) = 0$$

$$\tau_s \dot{\boldsymbol{\lambda}}_{a\mathrm{x}} = \left(\frac{\partial F}{\partial \mathbf{a}_\mathrm{x}}\right)^T \boldsymbol{\lambda}_c + \boldsymbol{\lambda}_{a\mathrm{x}} \qquad \text{with} \quad \boldsymbol{\lambda}_{a\mathrm{x}}(T) = 0$$

And gradient updates (**event-based**) can be written as:

$$\Delta w_{\mathrm{x},ij} = \frac{\partial}{\partial w_{\mathrm{x},ij}} \mathcal{L}(\mathbf{W}) = \sum_k \xi_{\mathrm{x},ijk}. \qquad \boldsymbol{\xi}_{\mathrm{x},k} = -\tau_s \left(\mathbf{r}_\mathrm{x}^-(s_k) \odot \mathbf{c}^-(s_k)\right) \otimes \boldsymbol{\lambda}_{a\mathrm{x}}^+(s_k)\,,$$

RUHR UNIVERSITÄT BOCHUM RUB

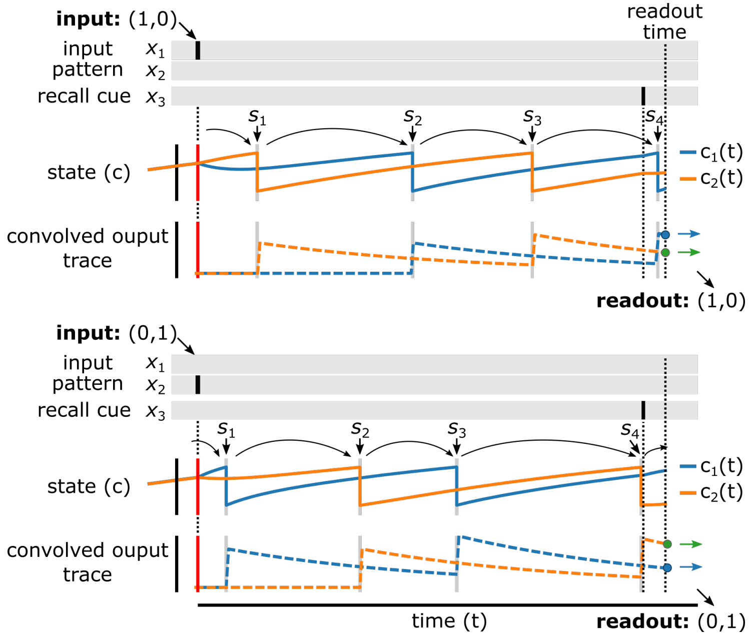
Preliminary results: Delay-copy

Delay copy task:

- Binary input pattern shown
- Output read out after recall cue

Network events convolved into trace for output

Trained with cross entropy loss to reach perfect recall



RUHR UNIVERSITÄT BOCHUM RUB

Summary

EGRU:

- Is a general event-based recurrent neural network architecture
- Exhibits high activity-sparsity as well as sparse learning updates
- Can be written in continuous time form that
  - supports event-based gradient descent updates
  - lends itself to rigorous mathematically analysis
- Can potentially replace SNNs for challenging and complex tasks

Outlook

- Explore other unit dynamics that is appropriate to different use-cases based on architectures that are
  - known to work well
  - a good fit for neuromorphic devices
    - E.g. non-binary packets for communication
- More efficient software implementations of such general event-based models
- Implementation on SpiNNaker 2 (and others?) and hopefully scale up to way more parameters and units

Joint work with

Khaleelulla Khan Nazeer
PhD student
TU Dresden

Mark Schöne
PhD student
TU Dresden

Christian Mayr
Professor
TU Dresden

David Kappel
Post-doc
RUB Bochum

Subramoney, A., Nazeer, K.K., Schöne, M., Mayr, C., Kappel, D., 2022.
EGRU: Event-based GRU for activity-sparse inference and learning.
https://doi.org/10.48550/arXiv.2206.06178

Thank you.
Questions?