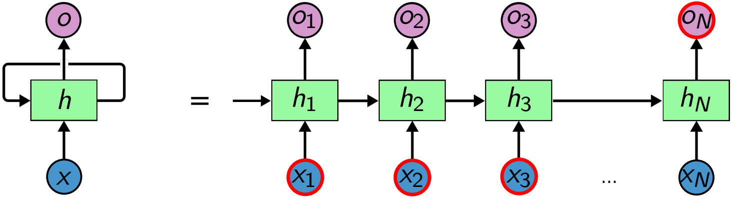
UnICORNN: A recurrent model for learning *very* long time dependencies

T. Konstantin Rusch   Siddhartha Mishra

Seminar for Applied Mathematics (SAM)
Department of Mathematics
ETH Zürich

Learning *very* long-term dependencies with RNNs

- Learning long-term dependencies with RNNs is difficult (Pascanu et al, 2013)
  - $\rightarrow$ mitigate exploding and vanishing gradient problem



- Learning very long-term dependencies with RNNs is very difficult
  - $\rightarrow$ mitigate exploding and vanishing gradient problem
  - $\rightarrow$ fast
  - $\rightarrow$ memory efficiency

UnICORNN architecture

- Base RNN on Hamiltonian system:

$$\boldsymbol{y}' = \boldsymbol{z}, \quad \boldsymbol{z}' = -[\sigma\left(\boldsymbol{w} \odot \boldsymbol{y} + \boldsymbol{V}\boldsymbol{u} + \boldsymbol{b}\right) + \alpha \boldsymbol{y}],$$

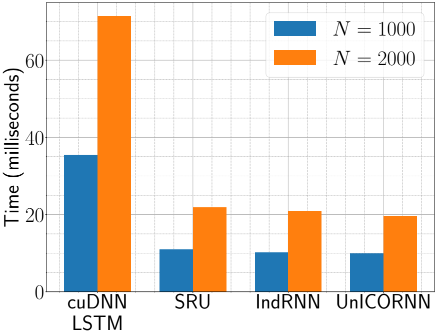
  hidden state $\boldsymbol{y}$, input $\boldsymbol{u}$.

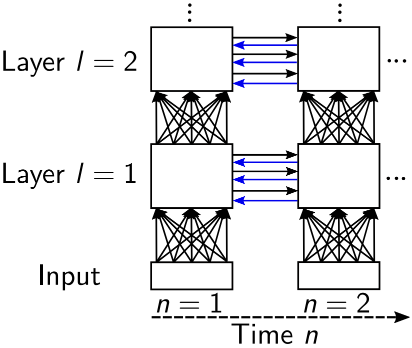
- Discretize with "learnable multi-scale symplectic Euler" and stack layers to obtain UnICORNN:

$$\begin{aligned}
\boldsymbol{y}_n^\ell &= \boldsymbol{y}_{n-1}^\ell + \Delta t \hat{\sigma}(\mathrm{c}^\ell) \odot \boldsymbol{z}_n^\ell, \\
\boldsymbol{z}_n^\ell &= \boldsymbol{z}_{n-1}^\ell - \Delta t \hat{\sigma}(\mathrm{c}^\ell) \odot [\sigma(\boldsymbol{w}^\ell \odot \boldsymbol{y}_{n-1}^\ell + \boldsymbol{V}^\ell \boldsymbol{y}_n^{\ell-1} + \boldsymbol{b}^l) + \alpha \boldsymbol{y}_{n-1}^\ell].
\end{aligned}$$

Properties of UnICORNN

- Gradients bounded → no exploding gradient
- Non-vanishing hidden state gradients → no vanishing gradient
- Invertible in time → memory efficient
- Multi-scale → increased expressivity
- Independent hidden states → very fast implementation on GPUs





Results

Table: Permuted sequential MNIST (seq. length = 784)

Model	test accuracy	# units	# params
LSTM	92.9%	256	270k
GRU	94.1%	256	200k
expRNN	96.6%	512	127k
coRNN	97.3%	256	134k
dense-IndRNN (L=6)	97.2%	128	257k
UnICORNN (L=3)	98.4%	256	135k

Table: Health-care: Vital sign prediction (seq. length = 4000).

Model	respiratory rate	heart rate
LSTM	2.28 ± 0.25	10.7 ± 2.0
expRNN	1.57 ± 0.16	1.87 ± 0.19
IndRNN (L=3)	1.47 ± 0.09	2.1 ± 0.2
coRNN	1.45 ± 0.23	1.71 ± 0.1
UnICORNN (L=3)	1.06 ± 0.03	1.39 ± 0.09

Results

Table: EigenWorms: Real-world (genomics) dataset (seq. length $\approx$ 18.000)

Model	test accuracy	# units	# params
t-BPTT LSTM	57.9% $\pm$ 7.0%	32	5.3k
sub-samp. LSTM	69.2% $\pm$ 8.3%	32	5.3k
expRNN	40.0% $\pm$ 10.1%	64	2.8k
IndRNN (L=2)	49.7% $\pm$ 4.8%	32	1.6k
coRNN	86.7% $\pm$ 3.0%	32	2.4k
UnICORNN (L=2)	90.3% $\pm$ 3.0%	32	1.5k

- We empirically show EigenWorms exhibits extreme long-term dependencies

Conclusion

- Propose new multi-layer recurrent model based on Hamiltonian system
  - No exploding/vanishing gradient problem
  - Multi-scale behavior
  - Memory efficient
  - Fast
- Achieve SOTA on many LTD benchmarks (length up to $\sim$ 18k)
- Set new high bar for very challenging real-world tasks
- UnICORNN based on very simple system: Only first step
- We will test on more real-world medical data

trusch@ethz.ch | github.com/tk-rusch/unicornn | @tk_rusch