Coupled Oscillatory Recurrent Neural Network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies

T. Konstantin Rusch Siddhartha Mishra

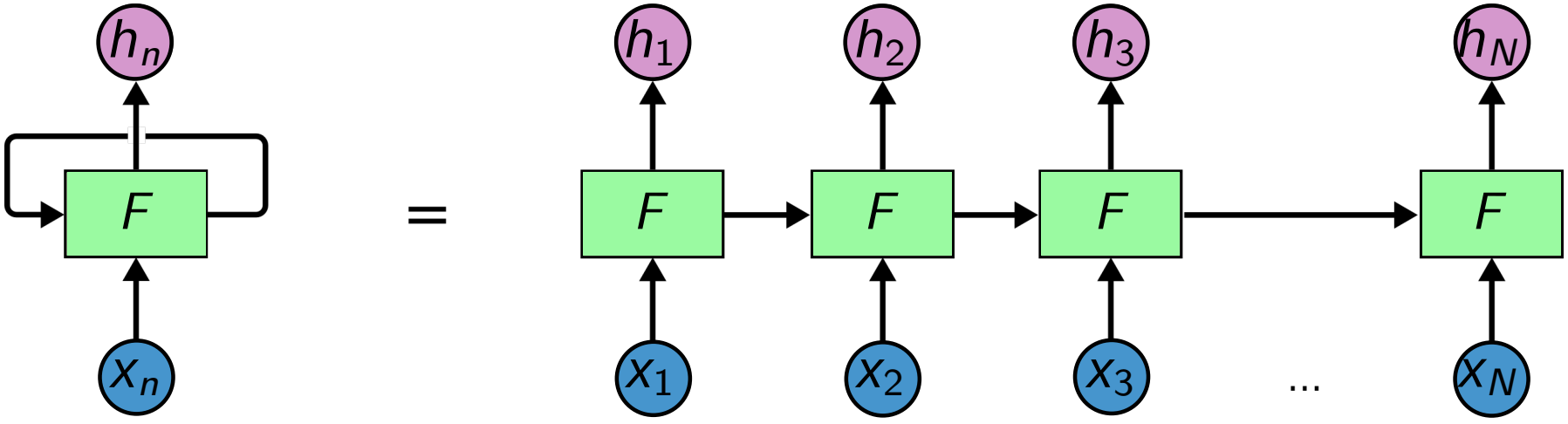
Seminar for Applied Mathematics (SAM)
Department of Mathematics
ETH Zürich

Recurrent neural networks

- RNNs achieved tremendous success in time series problems, e.g. in speech recognition, computer vision and natural language processing
- Recurrent Neural Network imposes temporal structure on network:

$$h_n = F(h_{n-1}, x_n, \Theta), \quad \forall n = 1, \ldots, N$$

- Training: Back-propagation through time
- Long term memory challenge



Difficulty of capturing long time dependencies

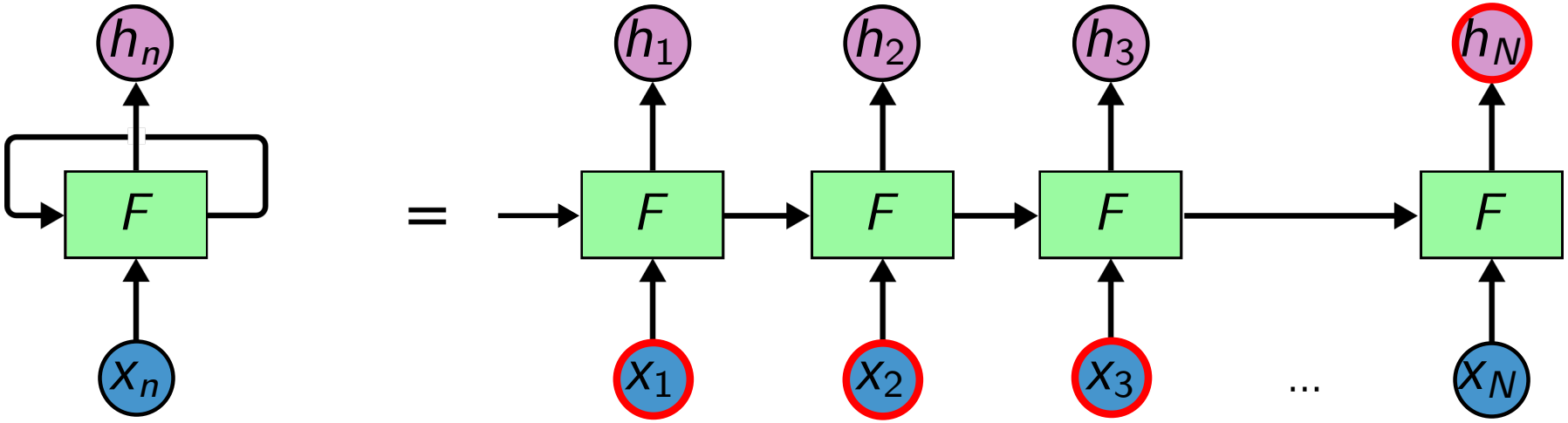
- Compute gradient of loss $\mathcal{E}$:

$$\frac{\partial \mathcal{E}}{\partial \theta} = \frac{1}{N} \sum_{n=1}^{N} \sum_{1 \leq k \leq n} \frac{\partial \mathcal{E}_n}{\partial h_n} \frac{\partial h_n}{\partial h_k} \frac{\partial^+ h_k}{\partial \theta}$$

- Possible exponential growth or decay of

$$\frac{\partial h_n}{\partial h_k} = \prod_{k < i \leq n} \frac{\partial h_i}{\partial h_{i-1}}$$

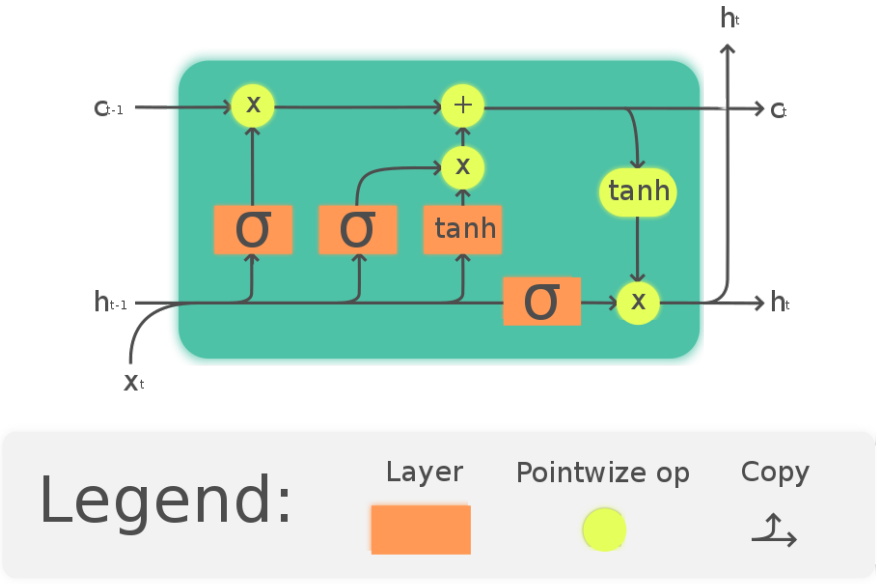
- Exploding/vanishing gradient problem (EVGP) (Pascanu et al, 2013)



Established solutions for learning LTDs

Gating mechanism:

- LSTM, GRU
- Additive gradient structure
- Gradients might still explode

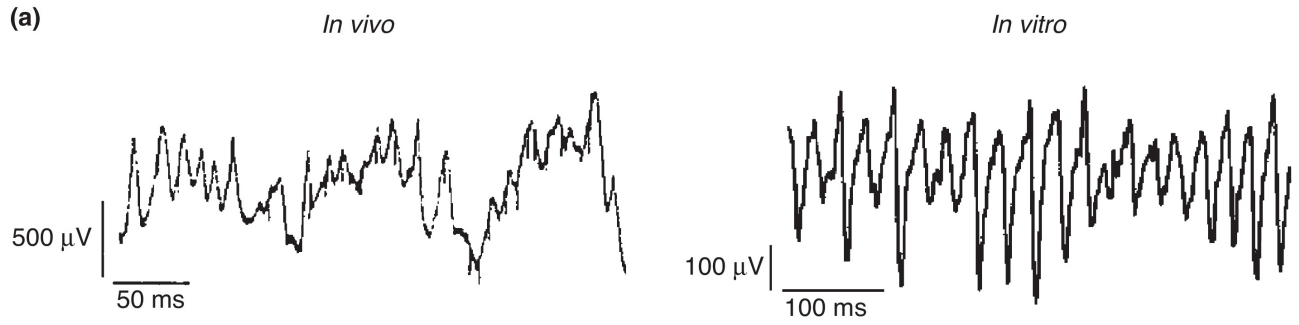


Constraining structure of hidden connection:

- unitary/orthogonal RNNs
- Enables RNNs to capture long time dependencies
- Structure constraints $\Rightarrow$ might lower expressivity

Coupled oscillators: A neurobiological inspiration

- Biological neurons viewed as oscillators
- Functional circuits of the brain interpreted by networks of oscillatory neurons



(Butler and Paulson, 2015)

- Abstract essence $\rightarrow$ RNNs based on coupled oscillators
- Stability of oscillatory dynamics: Expect EVGP is mitigated
- Coupled oscillators access very rich set of output states $\rightarrow$ high expressivity of system

Coupled oscillatory RNN (coRNN)

- Second-order system of ODEs:

$$y'' = \sigma\left(Wy + \mathcal{W}y' + Vu + b\right) - \gamma y - \epsilon y',$$

  hidden state y, input u.

- First order equivalent:

$$y' = z, \quad z' = \sigma\left(Wy + \mathcal{W}z + Vu + b\right) - \gamma y - \epsilon z.$$

- Discretize to obtain coupled oscillatory RNN (coRNN):

$$\begin{aligned} y_n &= y_{n-1} + \Delta t z_n, \\ z_n &= z_{n-1} + \Delta t \sigma\left(Wy_{n-1} + \mathcal{W}z_{n-1} + Vu_n + b\right) - \Delta t \gamma y_{n-1} - \Delta t \epsilon z_n. \end{aligned}$$

Exploding/vanishing gradient for coRNN

- Gradients for coRNN:

$$\frac{\partial \mathcal{E}}{\partial \theta} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \mathcal{E}_n}{\partial \theta} = \frac{1}{N}\sum_{n=1}^{N}\sum_{1\le k\le n}\underbrace{\frac{\partial \mathcal{E}_n}{\partial X_n}\frac{\partial X_n}{\partial X_k}\frac{\partial^+ X_k}{\partial \theta}}_{\frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta}}$$

- Assumption:

$$\max\left\{\frac{\Delta t(1+\|W\|_\infty)}{1+\Delta t}, \frac{\Delta t\|\mathcal{W}\|_\infty}{1+\Delta t}\right\} \le \Delta t^r, \quad \frac{1}{2} \le r \le 1.$$

Assumption easily met

Mitigation of EVGP for coRNN

- Exploding gradient problem for coRNN:

$$\left|\frac{\partial \mathcal{E}}{\partial \theta}\right| \leq \frac{3}{2}\left(m + \bar{Y}\sqrt{m}\right).$$

- Vanishing gradient problem for coRNN:

$$\frac{\partial \mathcal{E}_n^{(k)}}{\partial \theta} = \mathcal{O}\left(\hat{c}\delta\Delta t^{\frac{3}{2}}\right) + \mathcal{O}\left(\hat{c}\delta(1+\delta)\Delta t^{\frac{5}{2}}\right) + \mathcal{O}(\Delta t^3).$$

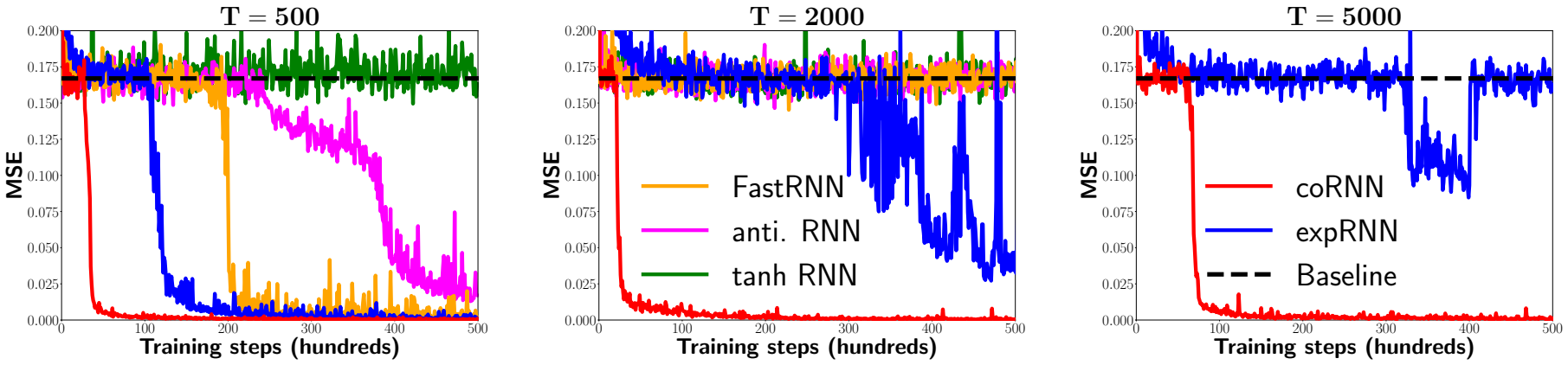
- Orders of estimate independent of k.

Experiment: Adding problem

- **Input**:

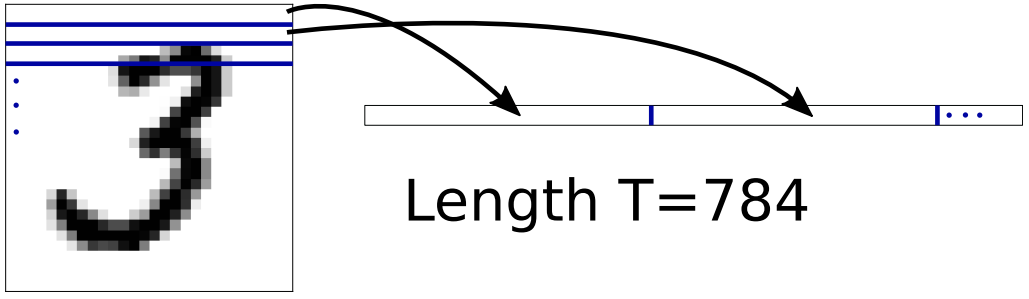
$$X = \begin{bmatrix} u_1 & \dots & u_i & \dots & u_{T/2} & \dots & u_j & \dots & u_T \\ 0 & \dots & 1 & \dots & 0 & \dots & 1 & \dots & 0 \end{bmatrix}, \quad u_t \sim \mathcal{U}([0,1])$$

- **Output**: $Y = u_i + u_j$
- **Goal:** Beat test MSE of 0.167 (variance of baseline output 1)



Experiment: (Permuted) Sequential MNIST

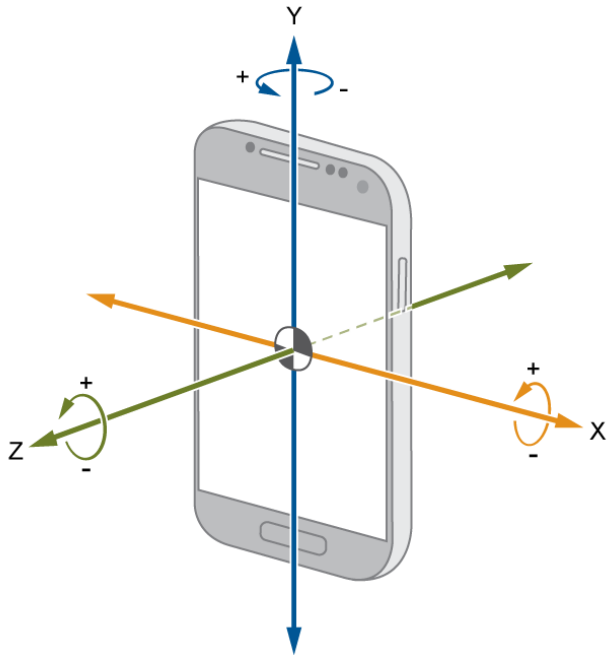
- sMNIST: Flatten MNIST images along rows
- psMNIST: Randomly permute sMNIST sequences



Model	sMNIST	psMNIST	# units	# params
LSTM	98.9%	92.9%	256	270k
GRU	99.1%	94.1%	256	200k
anti. RNN	98.0%	95.8%	128	10k
expRNN	98.7%	96.6%	512	137k
FastGRNN	98.7%	94.8%	128	18k
coRNN	99.3%	96.6%	128	34k
coRNN	99.4%	97.3%	256	134k

Experiment: Human activity recognition

- Collection of 6 tracked human activities, measured by accelerometer and gyroscope on smartphone
- **HAR-2**: Activities binarized to {Sitting, Laying, Walking_Upstairs} and {Standing, Walking, Walking_Downstairs}



Model	test accuracy	# units	# params
GRU	93.6%	75	19k
LSTM	93.7%	64	16k
FastGRNN	95.6%	80	7k
anti.sym. RNN	93.2%	120	8k
incremental RNN	96.3%	64	4k
coRNN	97.2%	64	9k
tiny coRNN	96.5%	20	1k

Experiment: IMDB sentiment analysis

- Collection of 50k movie reviews
- Sentiment analysis
- Tests expressivity of RNN

Model	test accuracy	# units	# params
LSTM	86.8%	128	220k
Skip LSTM	86.6%	128	220k
GRU	86.2%	128	164k
Skip GRU	86.6%	128	164k
ReLU GRU	84.8%	128	99k
expRNN	84.3%	256	58k
coRNN	87.4%	128	46k

Conclusion

- Neurobiologically inspired RNN
- Exploding and vanishing gradient problem mitigated
- SOTA results on LTD tasks
- coRNN as a first attempt $\rightarrow$ many possible extensions
- Plan to test coRNN on more oscillatory data

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