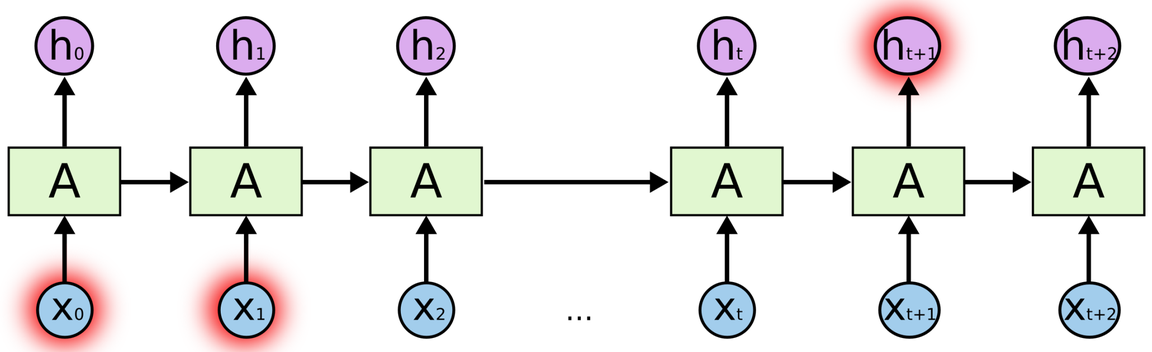
Gated Orthogonal Recurrent Units: On Learning to Forget

Li Jing, Çağlar Gülçehre, John Peurifoy, Yichen Shen,
Max Tegmark, Marin Soljačić, Yoshua Bengio

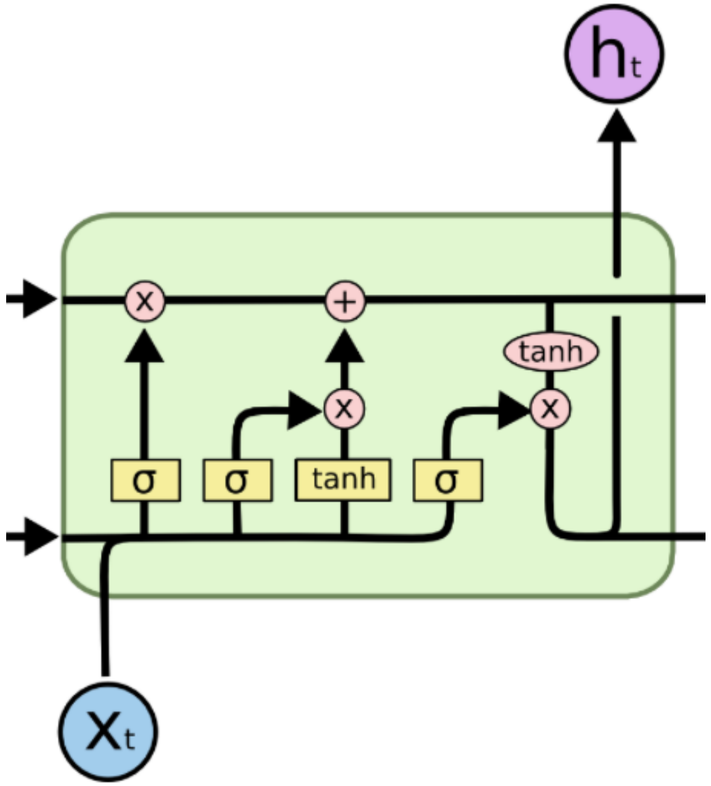
Massachusetts
Institute of
Technology

Gradient Vanishing/Explosion Problem



- During backpropagation through time, hidden to hidden Jacobian matrix is multiplied multiple times.
- Gradient vanishing/explosion makes RNN hard to train

Conventional Solution: LSTM



- Practically, gradient clipping is required
- slow to learn long term dependency

Unitary/Orthogonal RNN

Unitary/Orthogonal matrices keep the norm of vectors: $\|\boldsymbol{U}\boldsymbol{x}\| = \|\boldsymbol{x}\|$

By enforcing hidden to hidden transition matrix to be unitary/ orthogonal, no matter how many time steps are propagated, the norm of the gradient will stay the same

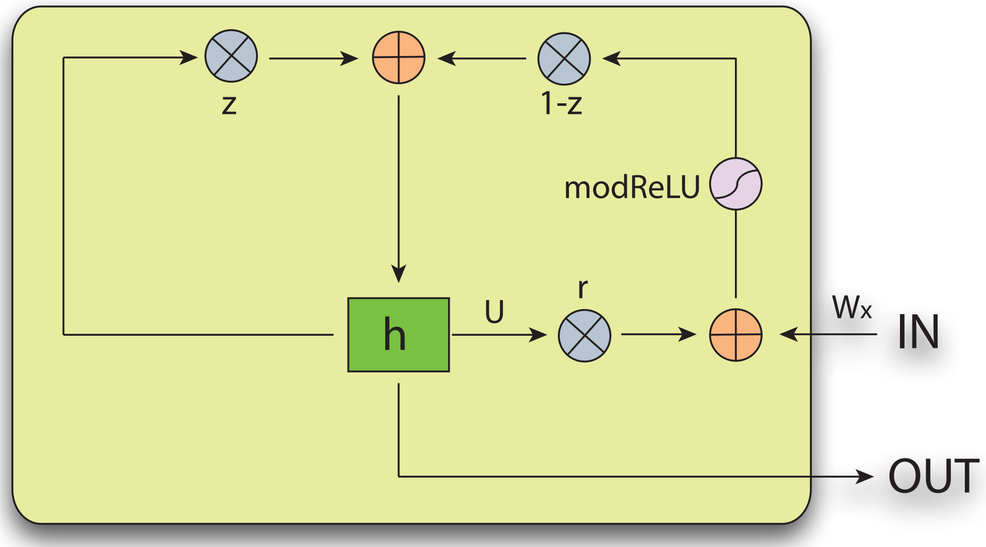
$$\left\| \prod_{k=t}^{T-1} \frac{\partial \mathbf{h}^{(k+1)}}{\partial \mathbf{h}^{(k)}} \right\| \sim 1$$

- Restricted-capacity Unitary Matrix Parametrization (Arjovsky, ICML 2016)
- Full-capacity Unitary Matrix by projection (Wisdom, NIPS 2016)
- Tunable Efficient Unitary Neural Networks (EUNN) and their application to RNN (Jing, ICML 2017)
- Orthogonal Matrix Parametrization by reflection (Mhammedi, ICML 2017)
- Orthogonal Matrix by regularization (Vorontsov, ICML 2017)

Limitation for basic Orthogonal RNN

- No forgetting mechanism
- Limited Memory size

Applying Gated System to Orthogonal RNN



Gated Orthogonal Recurrent Unit

Unitary/Orthogonal Matrices ——— Long Term Dependency

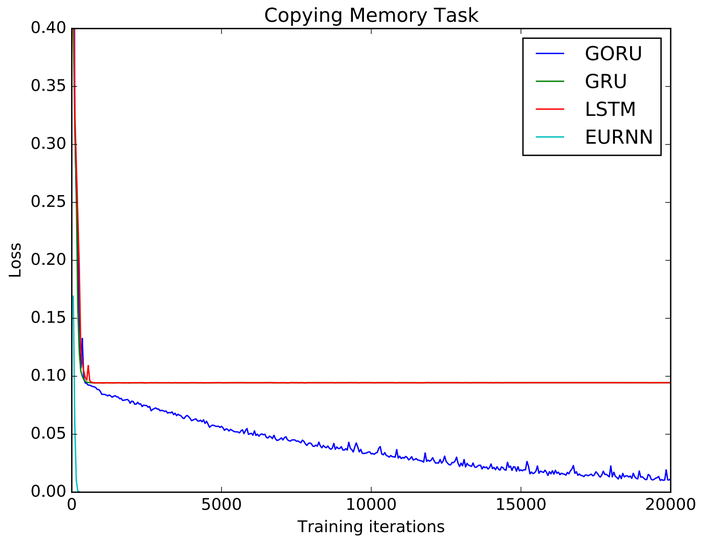
Gated Mechanism ——— Forgetting

Experiment results

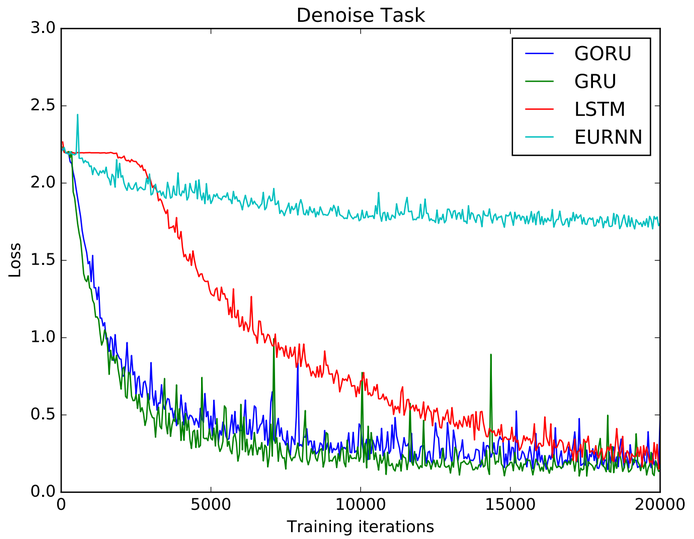
Synthetic Tasks:

GORU is the only one succeeding in all tasks

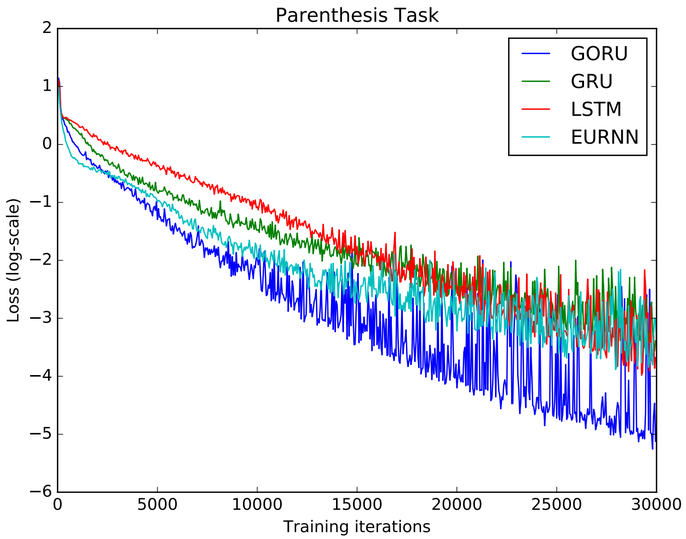
- Copying Task



- Denoise Task



- Parenthesis Task



Experiment results

Real Tasks: GORU outperforms all other models

- Speech Task

Model	#parameters	MSE(validation)	MSE(test)
LSTM	98k	58.8	57.5
GRU	72k	58.9	57.3
EURNN	41k	51.8	51.9
GORU	59k	**45.4**	**47.6**

- Question Answering Task

Task	GORU	GRU	LSTM	EURNN	baseline (Weston et al.)
1 - Single Supporting Fact	45.8	49.1	49.3	47.2	50
2 - Two Supporting Facts	39.5	38.5	32.3	24.3	20
3 - Three Supporting Facts	33.5	32.2	20.6	22.5	20
4 - Two Arg. Relations	62.7	64.6	67.5	56.1	61
5 - Three Arg. Relations	**87.0**	78.0	52.3	56.2	70
6 - Yes/No Questions	53.6	50.5	49.3	50.5	48
7 - Counting	77.7	79.5	76.9	71.9	49
8 - Lists/Sets	75.0	75.5	76.8	56.5	45
9 - Simple Negation	62.9	63.9	63.5	60.6	64
10 - Indefinite Knowledge	45.4	44.8	46.0	42.6	44
11 - Basic Coreference	69.3	71.2	71.1	72.1	72
12 - Conjunction	69.9	71.6	71.9	72.7	74
13 - Compound Coref.	92.7	94.2	93.8	92.4	94
14 - Time Reasoning	37.9	39.2	34.4	20.0	27
15 - Basic Deduction	55.2	57.4	20.9	25.0	21
16 - Basic Induction	44.0	45.9	45.9	43.3	23
17 - Positional Reasoning	**59.6**	50.5	51.6	51.2	51
18 - Size Reasoning	90.5	89.9	91.8	89.7	52
19 - Path Finding	8.9	9.6	8.2	9.0	8.0
20 - Agent's Motivations	97.7	97.7	96.5	93.3	91
Mean Performance	**60.4**	58.2	56.0	52.9	49.2

Thank you