Long-Short Term Memory

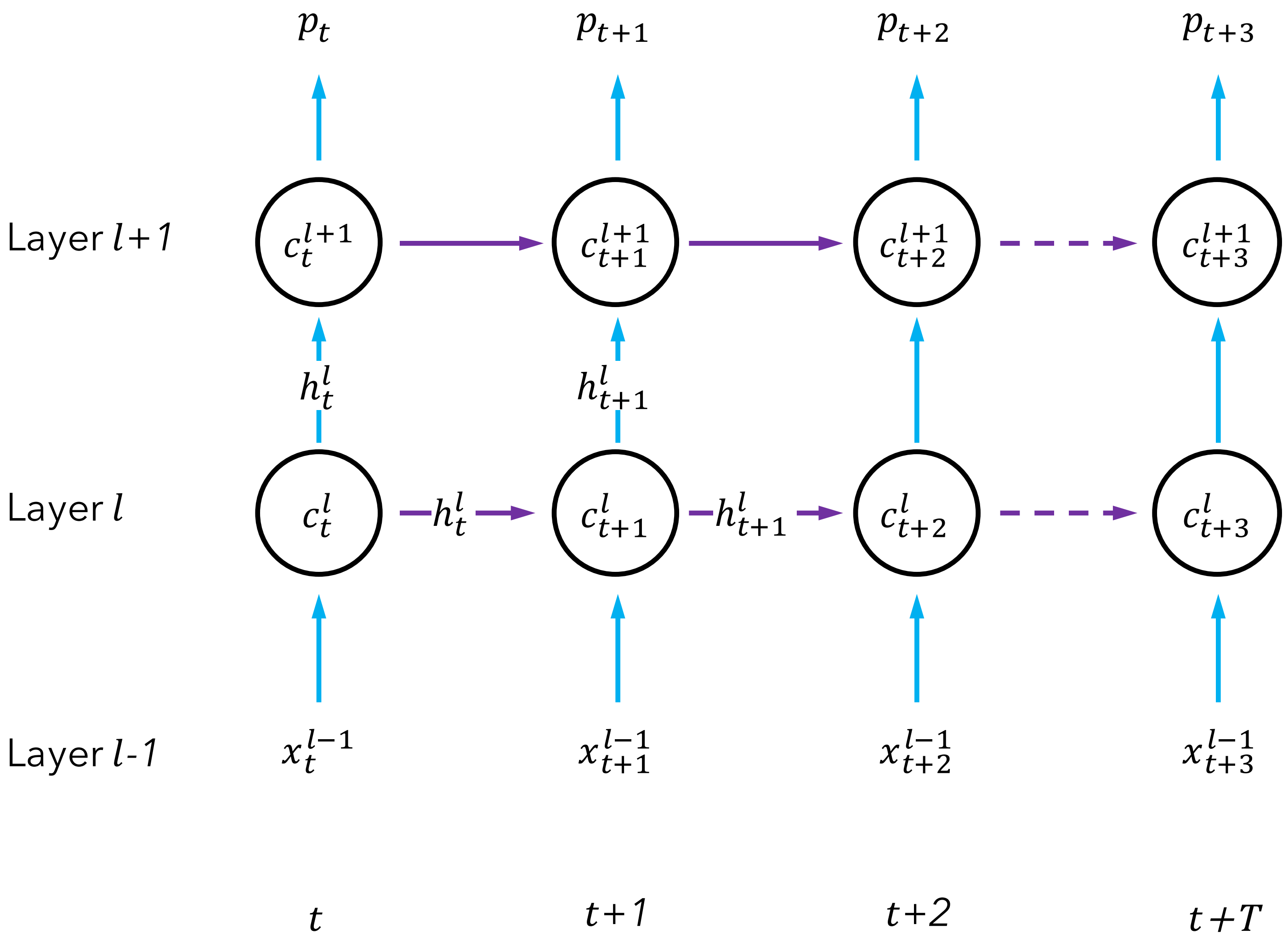
Authors: Hochreiter Sepp, Jürgen Schmidhuber

Presented by: Zihang Dai

Outline

- Motivation
- LSTM
- Experiments

Recurrent Neural Network

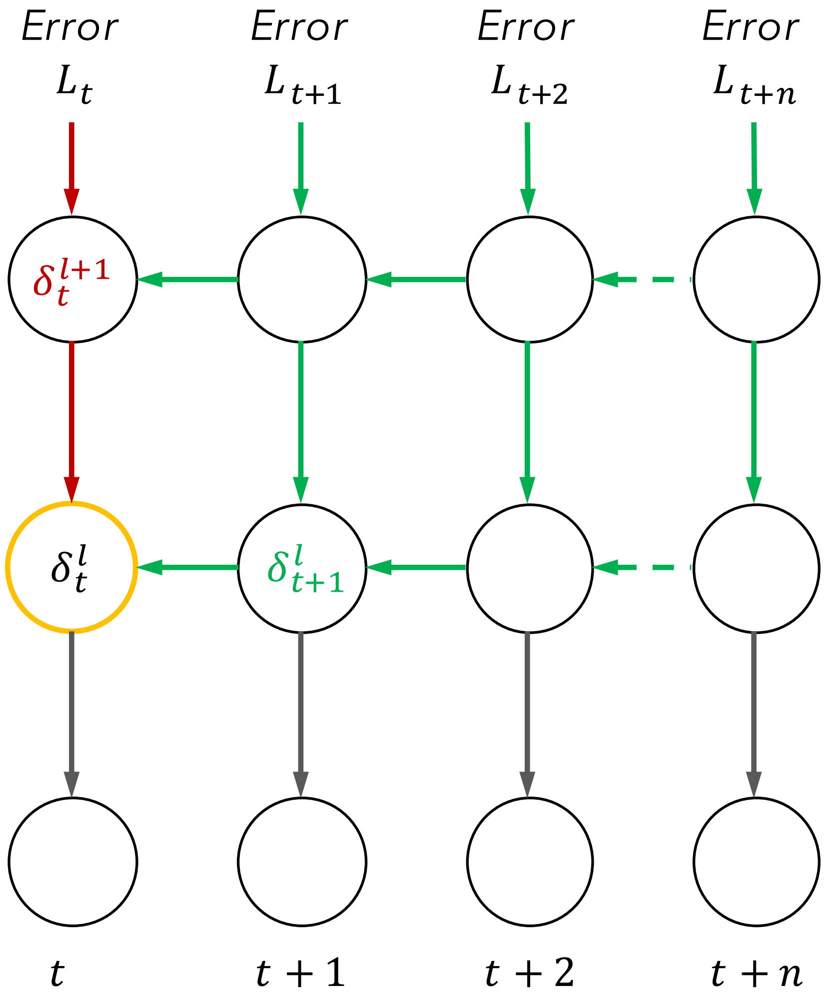


$$c_{t+1}^l = \boldsymbol{W_v} x_{t+1}^l + \boldsymbol{W_h} h_t^l + b$$

$$h_{t+1}^l = sigmoid\left(c_{t+1}^l\right)$$

$$L = \sum_{j=t}^{T} L_j = \sum_{j=t}^{T} \left(y_j - p_j\right)^2$$

Back-Propagation Through Time (BPTT)

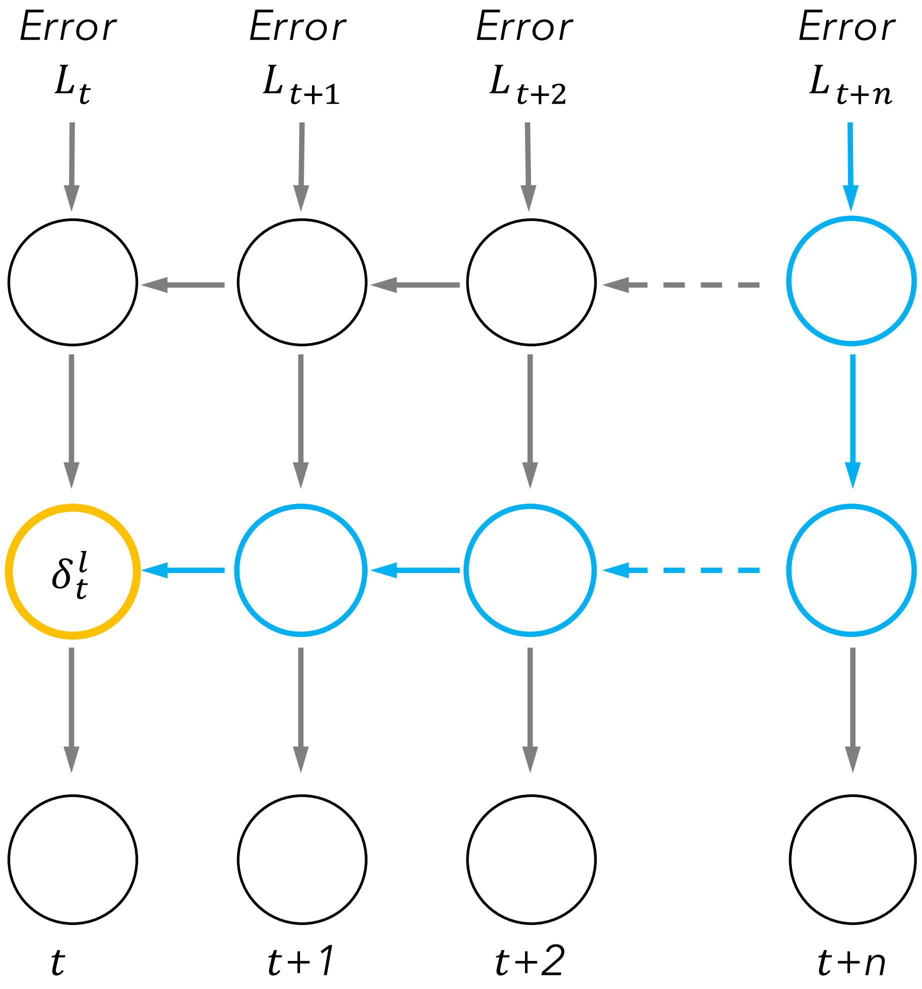


$$\delta_t^l = \frac{\partial}{\partial c_t^l}\sum_{j=t}^{T} L_j = \underbrace{\frac{\partial L_t}{\partial c_t^l}}_{\text{Spatial term}} + \overbrace{\frac{\partial}{\partial c_t^l}\sum_{j=t+1}^{T} L_j}^{\text{Temporal term}}$$

Spatial term: $$\frac{\partial L_t}{\partial c_t^{l+1}} \cdot \frac{\partial c_t^{l+1}}{\partial c_t^l} = \boldsymbol{\delta_t^{l+1}} \cdot \frac{\partial c_t^{l+1}}{\partial c_t^l}$$

Temporal term: $$\frac{\partial \sum_{j=t+1}^{T} L_j}{\partial c_{t+1}^l} \cdot \frac{\partial c_{t+1}^l}{\partial c_t^l} = \boldsymbol{\delta_{t+1}^l} \cdot \frac{\partial c_{t+1}^l}{\partial c_t^l}$$

Vanishing Gradient Problem



$$\frac{\partial L_{t+n}}{\partial c_t^l} = \frac{\partial L_{t+n}}{\partial c_{t+n}^l} \cdot \frac{\partial c_{t+n}^l}{\partial c_{t+n-1}^l} \cdot \dots \cdot \frac{\partial c_{t+1}^l}{\partial c_t^l}$$

$$= \frac{\partial L(t)}{\partial c_{t+n}^l} \cdot \underbrace{\prod_{\tau=t}^{t+n} \frac{\boldsymbol{\partial c_{\tau+1}^l}}{\boldsymbol{\partial c_{\tau}^l}}}_{\text{Sequential Jacobian}}$$

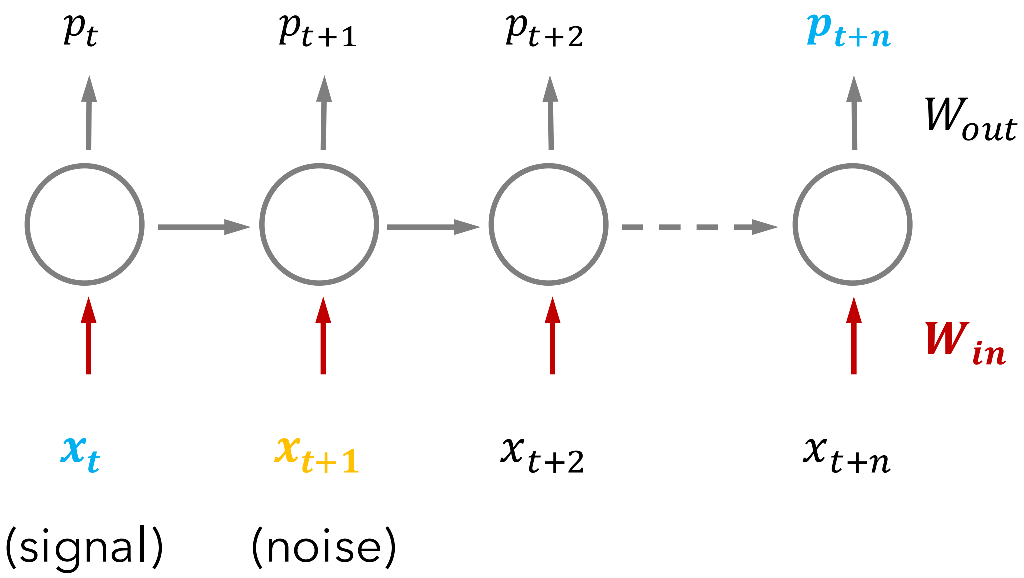
$$\frac{\boldsymbol{\partial c_{\tau+1}^l}}{\boldsymbol{\partial c_{\tau}^l}} = W_h^T \sigma'(c_\tau^l) \Rightarrow \left\| \frac{\boldsymbol{\partial c_{\tau+1}^l}}{\boldsymbol{\partial c_{\tau}^l}} \right\| \leq \|W_h\| \underbrace{\|\sigma'(c_\tau^l)\|}_{\leq 1/4}$$

- **Exponential** decayed error message
- **Long-term dependency** cannot be learned

Weight Conflict Problem

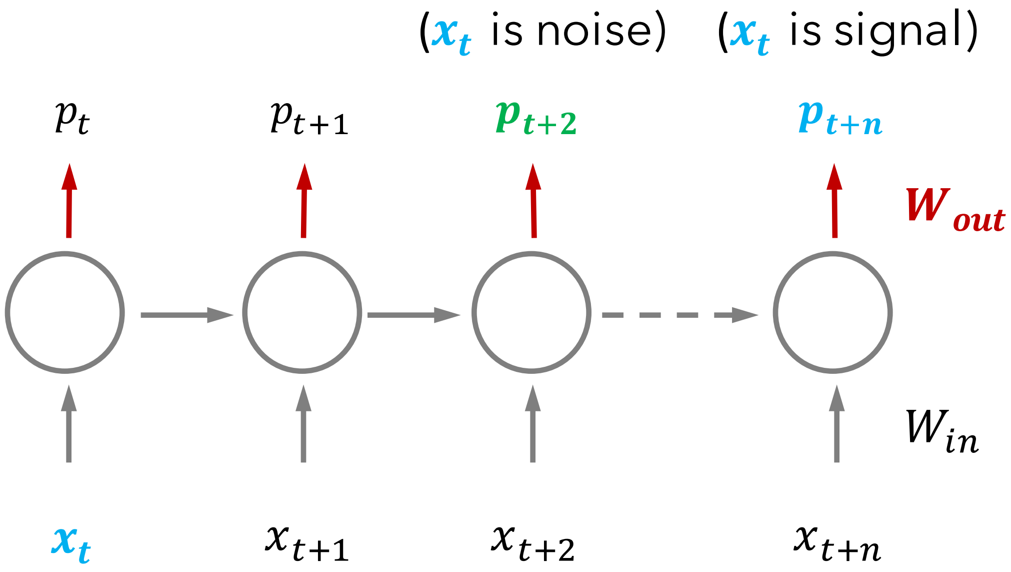
Two **conflict** roles of W_{in}

- **Absorb** useful signal x_t
- **Reject** harmful noise x_{t+1}



Two **conflict** roles of W_{out}

- **Reject** useless memory of x_t for p_{t+2}
- **Retrieve** useful memory of x_t for p_{t+n}



Treating Vanishing Gradient: Constant Error Carrousel (CEC)

Error signal doesn't vanish $\xRightarrow{\textit{i.e.}}$

$$\frac{\partial c^{l}_{\tau+1}}{\partial c^{l}_{\tau}} = W_h^T \sigma'\left(c^{l}_{\tau}\right) \boxed{\approx \boldsymbol{I}} \longrightarrow \textbf{CEC}$$

$\xRightarrow{\textit{e.g.}}$

Let $W = I, f\left(c^{l}_{\tau}\right) = c^{l}_{\tau}$

Then, $W_h^T \sigma'\left(c^{l}_{\tau}\right) = \boldsymbol{I}$

Problem with this idea

- No **non-linearity** (network won't be powerful)

Treating Wight Conflict: Gating Function

Core Idea →

Learn

1. what to store in the memory
2. what to retrieve from the memory

→

Gated Input

$$in_t = f_t^{in} \otimes x_t$$

- f_t^{in} is the input gating function
- $[f_t^{in}]_i \in [0, 1]$ (each element within [0,1])

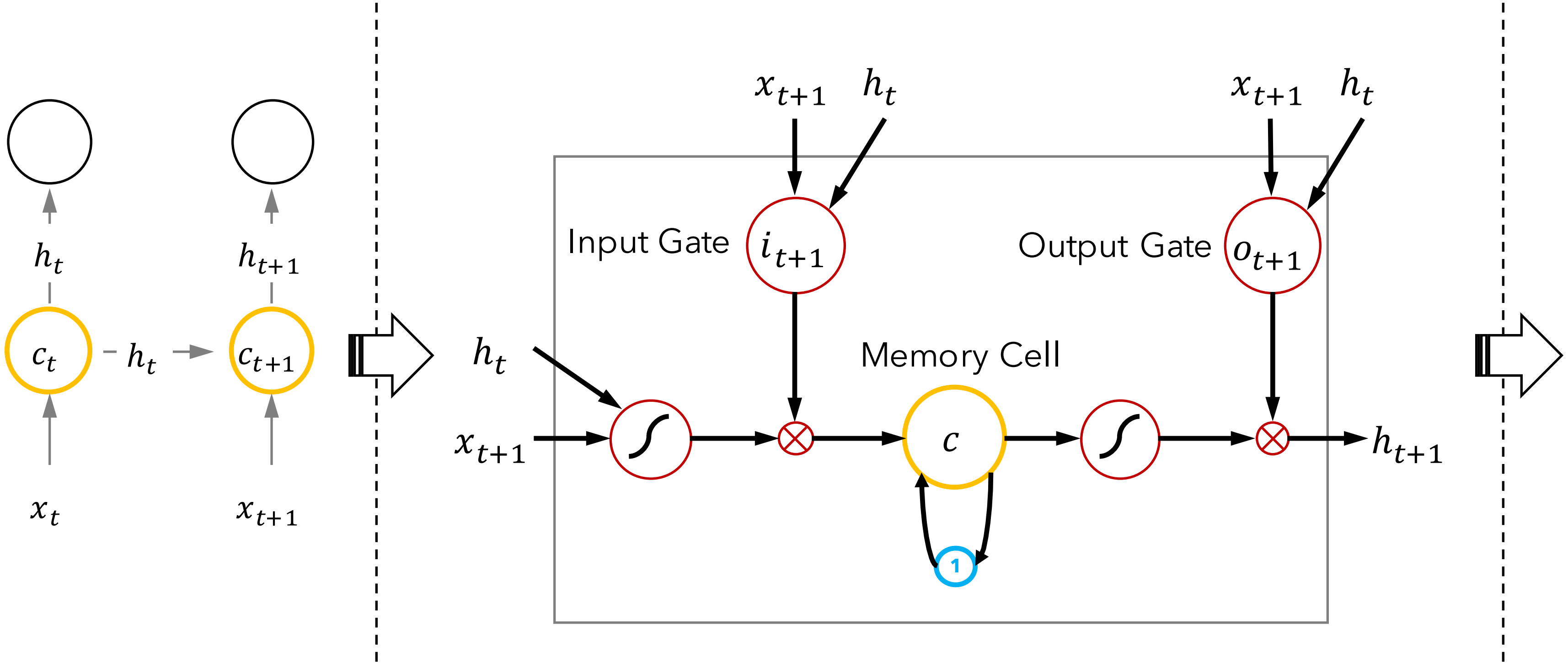
Gated Output

$$out_t = f_t^{out} \otimes c_t$$

- f_t^{out} is the output gating function
- $[f_t^{out}]_i \in [0, 1]$ (each element within [0,1])

($\otimes$ → element-wise multiplication)

CEC + Gates ➔ Long-Short Term Memory (LSTM)



Pure RNN

LSTM Cell Block

$$i_{t+1} = \sigma(M_{ix}x_{t+1} + M_{ih}h_t)$$

$$o_{t+1} = \sigma(M_{ox}x_{t+1} + M_{oh}h_t)$$

$$a_{t+1} = \phi(M_{cx}x_{t+1} + M_{ch}h_t)$$

$$c_{t+1} = c_t + i_{t+1} \otimes a_{t+1}$$

$$h_{t+1} = o_{t+1} \otimes \boldsymbol{\phi(c_{t+1})}$$

ϕ – *tanh* σ – *sigmoid*

① Constant Error Carrousel (CEC) ⊗ Element-wise multiplication

∫ Non-linearity i_{t+1} Input gate o_{t+1} Output gate

Why LSTM solves the problem

$$i_{t+1} = \sigma(M_{ix}x_{t+1} + M_{ih}h_t)$$

$$o_{t+1} = \sigma(M_{ox}x_{t+1} + M_{oh}h_t)$$

$$a_{t+1} = \phi(M_{cx}x_{t+1} + M_{ch}h_t)$$

$$c_{t+1} = c_t + i_{t+1} \otimes a_{t+1}$$

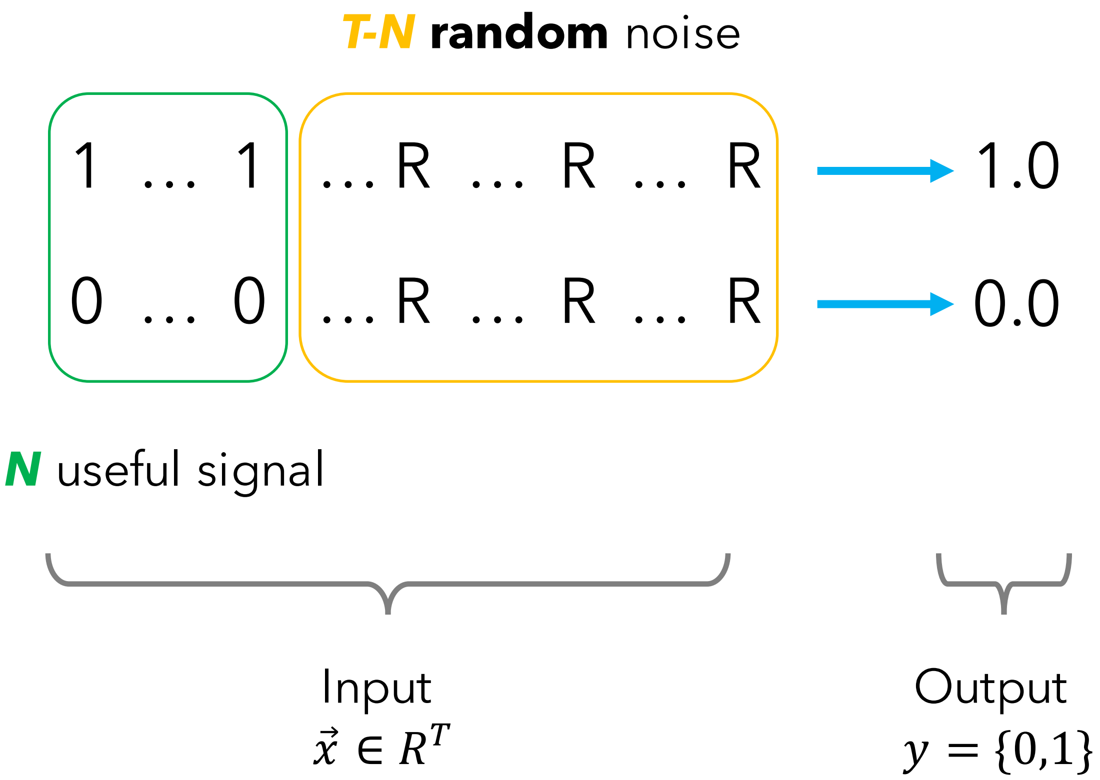
$$h_{t+1} = o_{t+1} \otimes \phi(c_{t+1})$$

$$\frac{\partial c_{\tau+1}^{l}}{\partial c_{\tau}^{l}} = I + \frac{\partial c_{\tau+1}^{l}}{\partial a_{\tau+1}^{l}} \cdot \frac{\partial a_{\tau+1}^{l}}{\partial h_{\tau}^{l}} \cdot \frac{\partial h_{\tau}^{l}}{\partial c_{\tau}^{l}} = \overbrace{\boldsymbol{I}}^{\textbf{CEC}} + \underbrace{\boldsymbol{i_{t+1}} \dots \boldsymbol{f_t} \dots}_{\textbf{Gated Error}}$$

$$\frac{\partial L_{t+n}}{\partial c_{t}^{l}} = \frac{\partial L_{t+n}}{\partial c_{t+n}^{l}} \prod_{\tau=t}^{t+n} \frac{\partial c_{\tau+1}^{l}}{\partial c_{\tau}^{l}} = \frac{\partial L_{t+n}}{\partial c_{t+n}^{l}} \prod_{j=1}^{n} (I + \ i_{t+1} \dots f_t \ \dots)$$

Experiment 3: two-sequence problem

Problem



T-N **random** noise

1 … 1 … R … R … R → 1.0

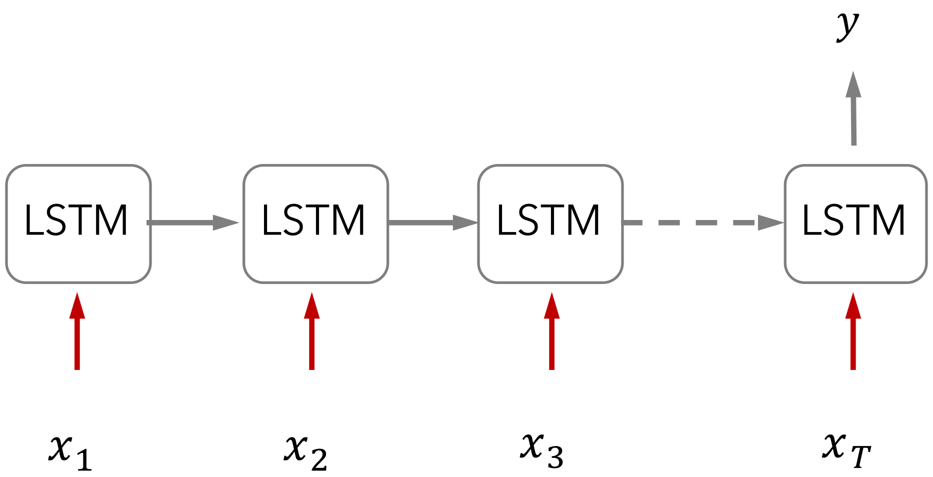
0 … 0 … R … R … R → 0.0

N useful signal

Input
$\vec{x} \in R^T$

Output
$y = \{0,1\}$

Model



Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult."

Experiment 4 & 5: adding/multiplication problem

Second dimension used as a marker

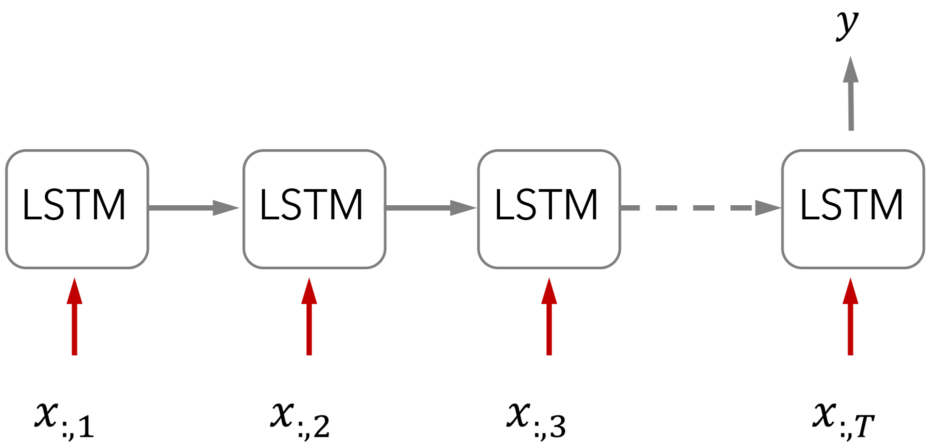
Adding Problem: $\cdots\begin{bmatrix}\mathbf{X_1}\\ \mathbf{-1}\end{bmatrix}\cdots\begin{bmatrix}R\\ 1\end{bmatrix}\cdots\begin{bmatrix}\mathbf{X_2}\\ \mathbf{-1}\end{bmatrix}\cdots\begin{bmatrix}R\\ 1\end{bmatrix}\cdots \longrightarrow 0.5+\frac{X_1+X_2}{4.0}$

Multiplication Problem: $\cdots\begin{bmatrix}\mathbf{X_1}\\ \mathbf{-1}\end{bmatrix}\cdots\begin{bmatrix}R\\ 1\end{bmatrix}\cdots\begin{bmatrix}\mathbf{X_2}\\ \mathbf{-1}\end{bmatrix}\cdots\begin{bmatrix}R\\ 1\end{bmatrix}\cdots \longrightarrow X_1 \times X_2$

Input: $X \in R^{2\times T}$

Output: $y \in R$

Model



Experiment 6: temporal order problem

Problem

… R … X … X … R … → Q

… R … X … Y … R … → R

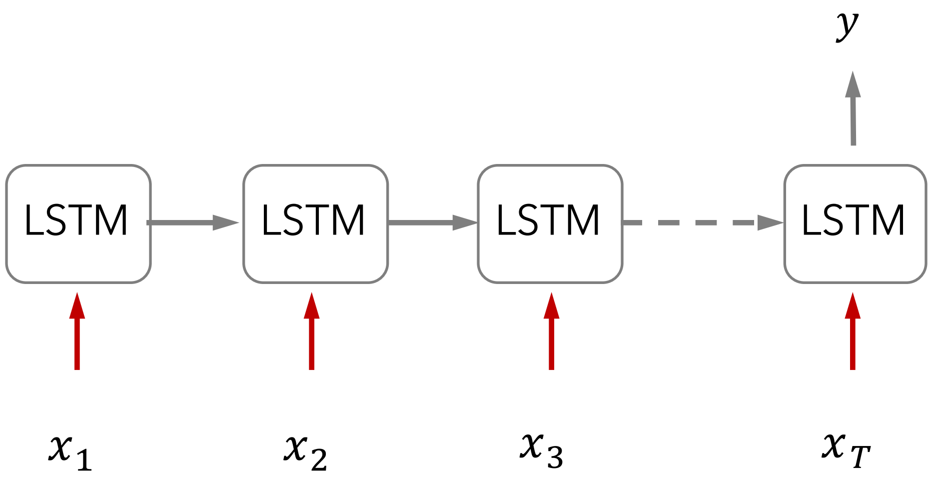
… R … Y … X … R … → S

… R … Y … Y … R … → U

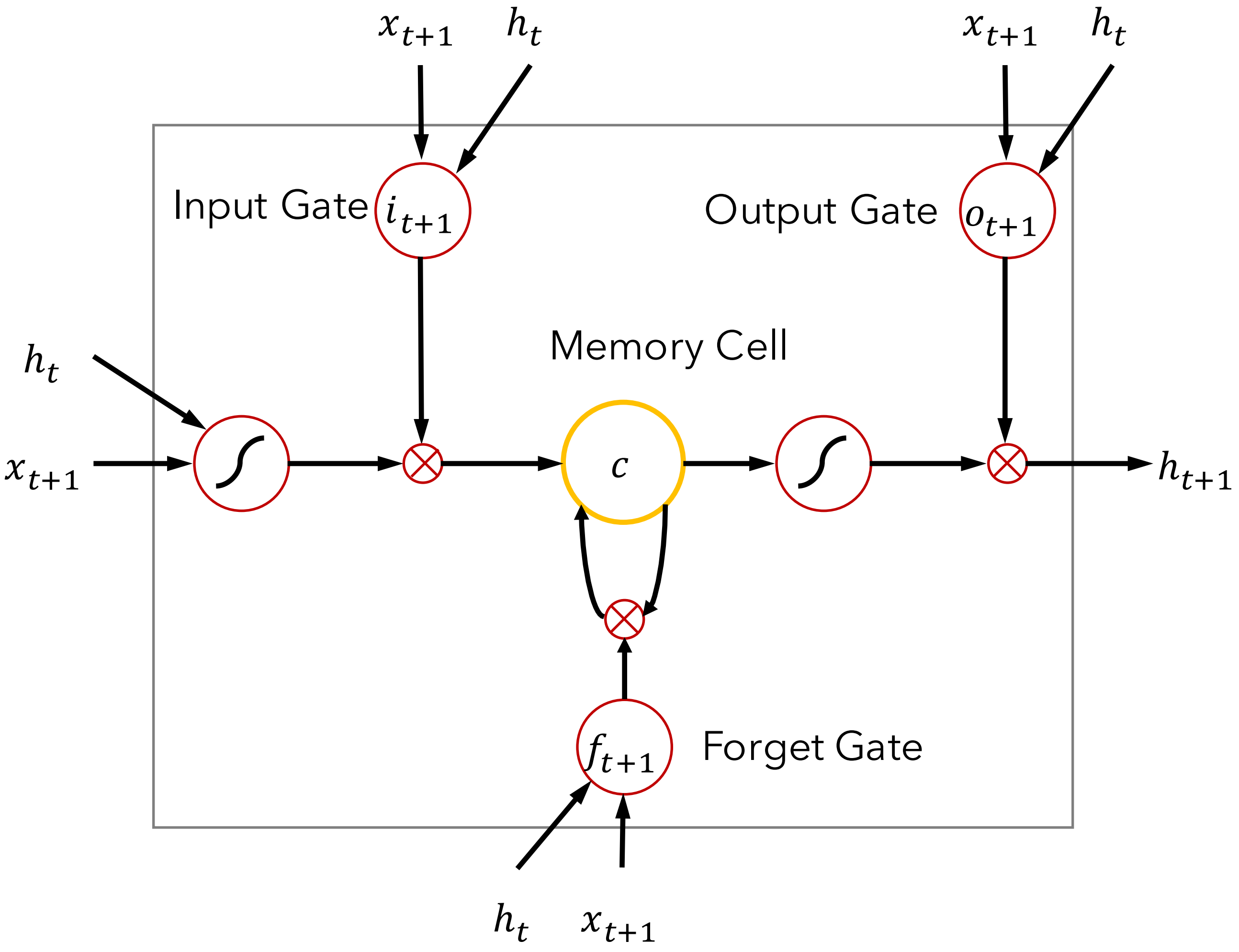
Input
$\vec{x} \in R^T$

Output
$y = \{Q, R, S, U\}$

Model



Introducing Forget Gate



$$i_{t+1} = \sigma(M_{ix}x_{t+1} + M_{ih}h_t)$$

$$f_{t+1} = \sigma(M_{fx}x_{t+1} + M_{fh}h_t)$$

$$o_{t+1} = \sigma(M_{ox}x_{t+1} + M_{oh}h_t)$$

$$a_{t+1} = \phi(M_{cx}x_{t+1} + M_{ch}h_t)$$

$$c_{t+1} = \boldsymbol{f_{t+1}} \otimes \boldsymbol{c_t} + i_{t+1} \otimes a_{t+1}$$

$$h_{t+1} = o_{t+1} \otimes \phi(c_{t+1})$$

Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM."

Thanks & Questions