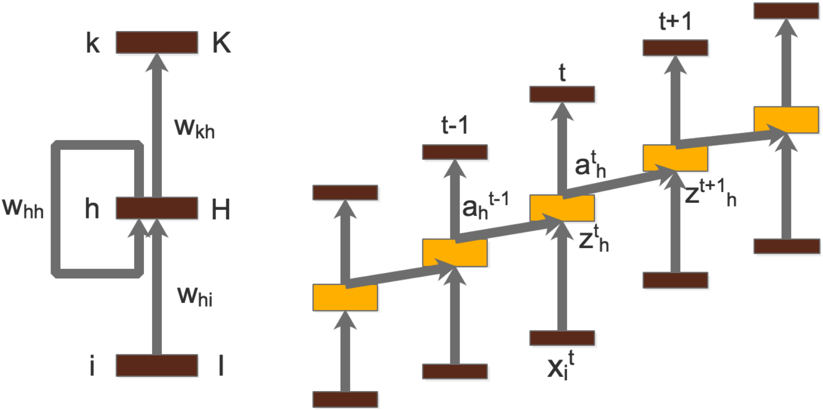
Deep Learning Technology and Application

Ge Li

Peking University

RNN 符号体系



RNN 前向传播推导

x: 长度为 T 的输入；x_i^t:t 时刻的输入 x 的第 i 维；
I: 输入层神经元个数；H: 隐藏层神经元个数；K: 输出层神经元个数；
z_j^t: 神经元 j 在 t 时刻的待激活输入；
a_j^t: 神经元 j 在 t 时刻的激活值；
J^t: 用 t 时刻的输出计算的代价函数；

$$z_h^t = \sum_{i=1}^{I} w_{hi} x_i^t + \sum_{h'=1}^{H} w_{hh'} a_{h'}^{t-1} \qquad a_h^t = f_h(z_h^t)$$

$$z_k^t = \sum_{h=1}^{H} w_{kh} a_h^t$$

$$J = \sum_{t=1}^{T} J^t(W, b)$$

其中，a_i^0 需要进行初始化，可以选择 0，也可以选择非零初始值。

RNN 后向传播推导

$$
\begin{aligned}
w_{kh} &= w_{kh} - \frac{\partial J}{\partial w_{kh}} \\
&= w_{kh} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial w_{kh}} \\
&= w_{kh} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial z_k^t} \frac{\partial z_k^t}{\partial w_{kh}}
\end{aligned}
$$

因为：$z_k^t = \sum_{h=1}^{H} w_{kh} a_h^t$ 所以：

$$
w_{kh} = w_{kh} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial z_k^t} a_h^t
$$

RNN 后向传播推导

$$w_{hh'} = w_{hh'} - \frac{\partial J}{\partial w_{hh'}} = w_{hh'} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial w_{hh'}}$$

$$= w_{hh'} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial z_h^t} \frac{\partial z_h^t}{\partial w_{hh'}}$$

因为：$z_h^t = \sum_{i=1}^{I} w_{hi} x_i^t + \sum_{h'=1}^{H} w_{hh'} a_{h'}^{t-1}$ 所以：

$$= w_{hh'} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial z_h^t} a_{h'}^{t-1}$$

设：$\delta_h^t = \frac{\partial J^t(W,b)}{\partial z_h^t}$ 则：$w_{hh'} = w_{hh'} - \sum_{t=1}^{T} \delta_h^t a_{h'}^{t-1}$

RNN 后向传播推导

$$w_{hi} = w_{hi} - \frac{\partial J}{\partial w_{hi}} = w_{hi} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial w_{hi}}$$

$$= w_{hi} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial z_h^t} \frac{\partial z_h^t}{\partial w_{hi}}$$

因为：$z_h^t = \sum_{i=1}^{I} w_{hi} x_i^t + \sum_{h'=1}^{H} w_{hh'} a_{h'}^{t-1}$ 所以：

$$= w_{hi} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial z_h^t} x_i^t$$

设：$\delta_h^t = \frac{\partial J^t(W,b)}{\partial z_h^t}$ 则：$w_{hi} = w_{hi} - \sum_{t=1}^{T} \delta_h^t x_i^t$

RNN 后向传播推导

因为：

$$z_h^t = \sum_{i=1}^{I} w_{hi} x_i^t + \sum_{h'=1}^{H} w_{hh'} a_{h'}^{t-1}$$

所以：

$$\begin{aligned}
\delta_h^t = \frac{\partial J(W,b)}{\partial z_h^t} &= \sum_{k=1}^{K} \frac{\partial J^t}{\partial z_k^t} \frac{\partial z_k^t}{\partial z_h^t} + \sum_{h=1}^{H} \frac{\partial J^{t+1}}{\partial z_h^{t+1}} \frac{\partial z_h^{t+1}}{\partial z_h^t} \\
&= \sum_{k=1}^{K} \frac{\partial J^t}{\partial z_k^t} \frac{\partial z_k^t}{\partial a_h^t} \frac{\partial a_h^t}{\partial z_h^t} + \sum_{h=1}^{H} \frac{\partial J^{t+1}}{\partial z_h^{t+1}} \frac{\partial z_h^{t+1}}{\partial a_{h'}^t} \frac{\partial a_{h'}^t}{\partial z_h^t} \\
&= \sum_{k=1}^{K} \delta_k^t w_{kh} f_h'(\cdot) + \sum_{h=1}^{H} \delta_h^{t+1} w_{hh'} f_h'(\cdot) \\
&= \left(\sum_{k=1}^{K} \delta_k^t w_{kh} + \sum_{h=1}^{H} \delta_h^{t+1} w_{hh'} \right) f_h'(\cdot)
\end{aligned}$$

RNN 后向传播推导

$$w_{hk} = w_{hk} - \sum_{t=1}^{T} \frac{\partial J^t(W,b)}{\partial z_k^t} a_h^t$$

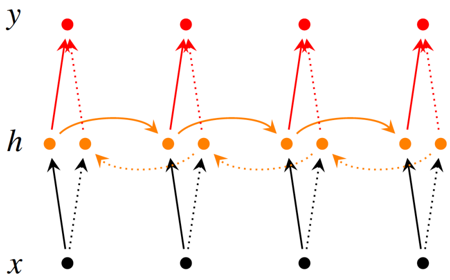
$$w_{hi} = w_{hi} - \sum_{t=1}^{T} \delta_h^t x_i^t$$

$$w_{h'h} = w_{h'h} - \sum_{t=1}^{T} \delta_h^t a_{h'}^{t-1}$$

$$\delta_h^t = \left(\sum_{k=1}^{K} \delta_k^t w_{kh} + \sum_{h'=1}^{H} \delta_h^{t+1} w_{hh'} \right) f_h'(\cdot)$$

Bi-directional RNN

Bi-directional RNN

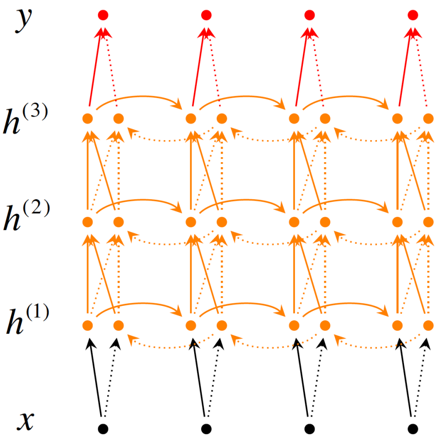


$$\overrightarrow{h}_t = f(\overrightarrow{W} x_t + \overrightarrow{V}\,\overrightarrow{h}_{t-1} + \overrightarrow{b})$$

$$\overleftarrow{h}_t = f(\overleftarrow{W} x_t + \overleftarrow{V}\,\overleftarrow{h}_{t+1} + \overleftarrow{b})$$

$$\hat{y}_t = g(Uh_t + c) = g(U[\overrightarrow{h}_t; \overleftarrow{h}_t] + c)$$

Bi-directional RNN

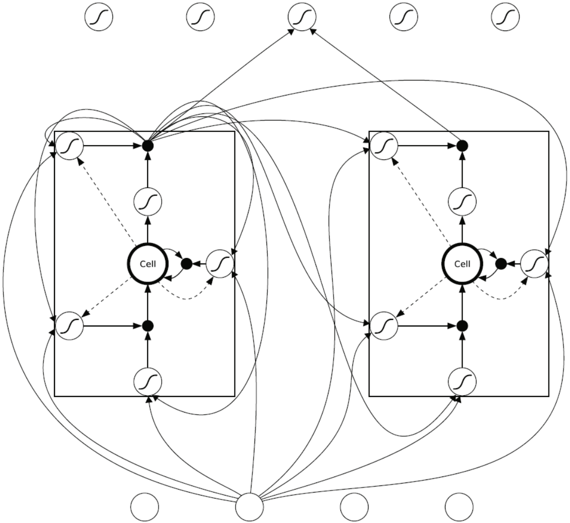


$$\overrightarrow{h}_t^{(i)} = f(\overrightarrow{W}^{(i)} h_t^{(i-1)} + \overrightarrow{V}^{(i)} \overrightarrow{h}_{t-1}^{(i)} + \overrightarrow{b}^{(i)})$$

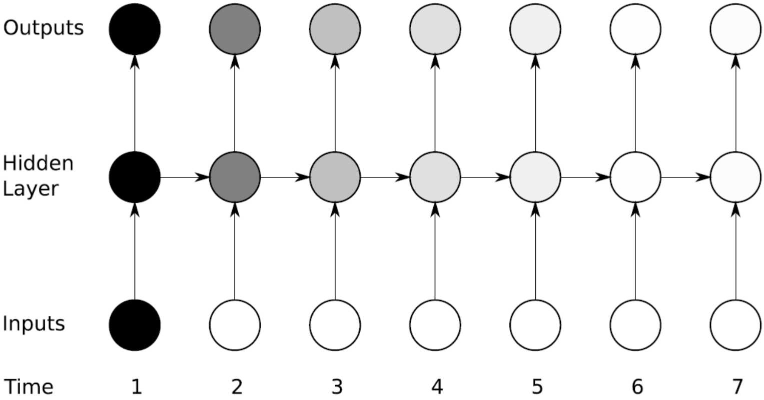
$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t+1}^{(i)} + \overleftarrow{b}^{(i)})$$

$$\hat{y}_t = g(Uh_t + c) = g(U[\overrightarrow{h}_t^{(L)}; \overleftarrow{h}_t^{(L)}] + c)$$

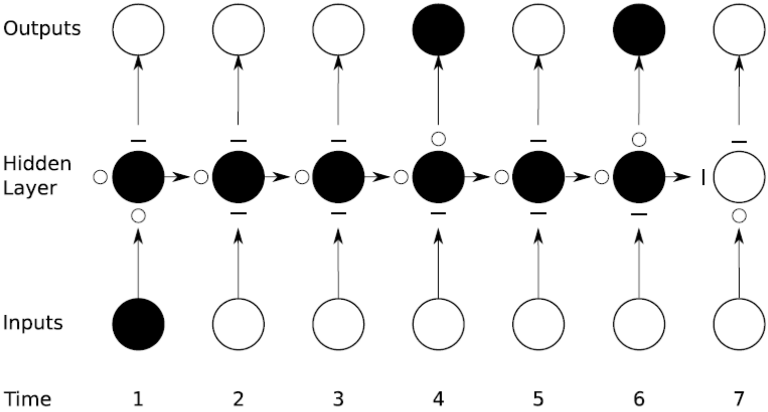
Recurrent Neural Network ——LSTM



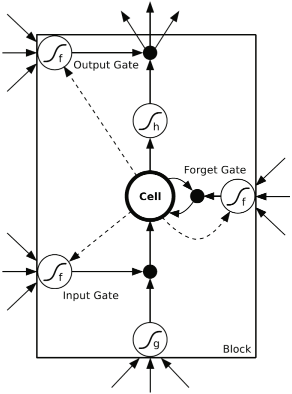
LSTM 的提出



LSTM 的提出

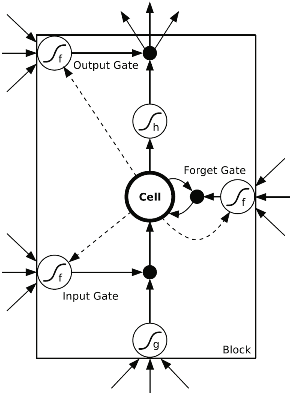


LSTM 单元



- 设 LSTM 隐藏层共包含 H 个神经元，下标 h 表示其中之一；
- 设 LSTM 隐藏层共包含 C 个 Cell，下标 c 表示某个 Cell；
- 当前的 LSTM 单元中 Input Gate,Forget Gate, Output Gate 分别用下标 α, β, γ 标识；
- 第 h 个 LSTM 单元在 t 时刻的输入：z_h^t， t 时刻的输出：a_h^t；
- 第 k 个输出层神经元单元在 t 时刻的输入：z_k^t， t 时刻的输出：a_k^t；

LSTM 单元

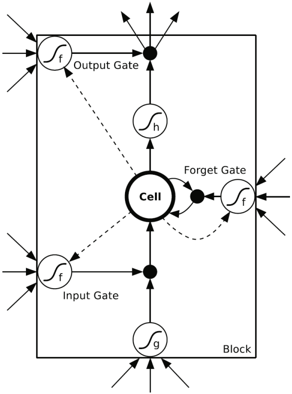


LSTM 单元的输入:

$$z_h^t = \sum_{i=1}^{I} w_{hi} x_i^t + \sum_{h=1}^{H} w_{hh} a_h^{t-1}$$

其中：a_h 表示来自于其他 LSTM 单元的输出；

LSTM 单元



Input Gate:

$$z_\alpha^t = \sum_{i=1}^{I} w_{\alpha i} x_i^t + \sum_{h=1}^{H} w_{\alpha h} a_h^{t-1} + \sum_{c=1}^{C} w_{\alpha c} s_c^{t-1}$$

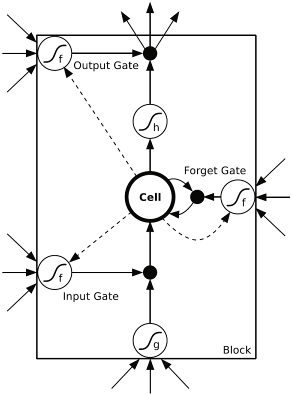
$$a_\alpha^t = f(z_\alpha^t)$$

Forget Gate:

$$z_\beta^t = \sum_{i=1}^{I} w_{\beta i} x_i^t + \sum_{h=1}^{H} w_{\beta h} a_h^{t-1} + \sum_{c=1}^{C} w_{\beta c} s_c^{t-1}$$

$$a_\beta^t = f(z_\beta^t)$$

LSTM 单元

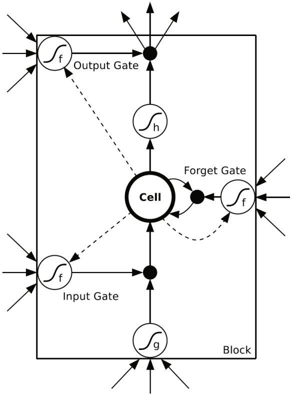


Output Gate:

$$z_\gamma^t = \sum_{i=1}^{I} w_{\gamma i} x_i^t + \sum_{h=1}^{H} w_{\gamma h} a_h^{t-1} + \sum_{c=1}^{C} w_{\gamma c} s_c^{t-1}$$

$$a_\gamma^t = f(z_\gamma^t)$$

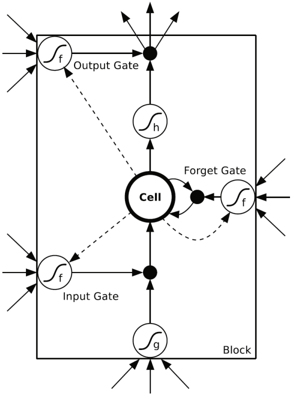
LSTM 单元



Cells:

$$s_c^t = a_\alpha^t g(z_h^t) + a_\beta^t s_c^{t-1}$$

LSTM 单元



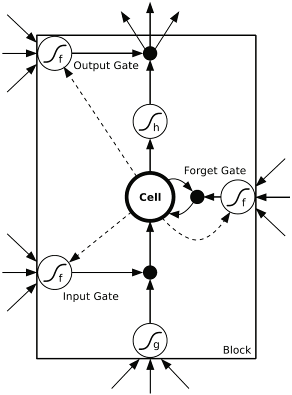
LSTM Cell Outputs:

$$a_h^t = a_\gamma^t h(s_c^t)$$

Cell in RNN output layer:

$$z_k^t = \sum_{h=1}^{H} w_{kh} a_h^t$$

LSTM 单元



统计要计算的参数：

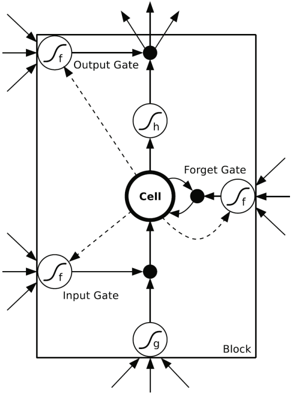
$$w_{\alpha i}, w_{\alpha h}, w_{\alpha c}$$

$$w_{\beta i}, w_{\beta h}, w_{\beta c}$$

$$w_{\gamma i}, w_{\gamma h}, w_{\gamma c}$$

$$w_{hi}, w_{hh}, w_{kh}$$

LSTM 单元



$$w_{\alpha i}: \frac{\partial J(W,b)}{\partial w_{\alpha i}} = \frac{\partial J(W,b)}{\partial z_\alpha^t} \frac{\partial z_\alpha^t}{\partial w_{\alpha i}} = \frac{\partial J(W,b)}{\partial z_\alpha^t} x_i^t$$

$$w_{\alpha h}: \frac{\partial J(W,b)}{\partial w_{\alpha h}} = \frac{\partial J(W,b)}{\partial z_\alpha^t} \frac{\partial z_\alpha^t}{\partial w_{\alpha h}} = \frac{\partial J(W,b)}{\partial z_\alpha^t} a_h^{t-1}$$

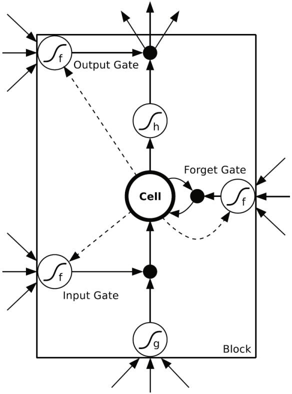
$$w_{\alpha c}: \frac{\partial J(W,b)}{\partial w_{\alpha c}} = \frac{\partial J(W,b)}{\partial z_\alpha^t} \frac{\partial z_\alpha^t}{\partial w_{\alpha c}} = \frac{\partial J(W,b)}{\partial z_\alpha^t} s_c^{t-1}$$

$$w_{\beta i}: \frac{\partial J(W,b)}{\partial w_{\beta i}} = \frac{\partial J(W,b)}{\partial z_\beta^t} \frac{\partial z_\beta^t}{\partial w_{\beta i}} = \frac{\partial J(W,b)}{\partial z_\beta^t} x_i^t$$

$$w_{\beta h}: \frac{\partial J(W,b)}{\partial w_{\beta h}} = \frac{\partial J(W,b)}{\partial z_\beta^t} \frac{\partial z_\beta^t}{\partial w_{\beta h}} = \frac{\partial J(W,b)}{\partial z_\beta^t} a_h^{t-1}$$

$$w_{\beta c}: \frac{\partial J(W,b)}{\partial w_{\beta c}} = \frac{\partial J(W,b)}{\partial z_\beta^t} \frac{\partial z_\beta^t}{\partial w_{\beta c}} = \frac{\partial J(W,b)}{\partial z_\beta^t} s_c^{t-1}$$

LSTM 单元



$$w_{\gamma i}: \frac{\partial J(W,b)}{\partial w_{\gamma i}} = \frac{\partial J(W,b)}{\partial z_\gamma^t}\frac{\partial z_\gamma^t}{\partial w_{\gamma i}} = \frac{\partial J(W,b)}{\partial z_\gamma^t} x_i^t$$

$$w_{\gamma h}: \frac{\partial J(W,b)}{\partial w_{\gamma h}} = \frac{\partial J(W,b)}{\partial z_\gamma^t}\frac{\partial z_\gamma^t}{\partial w_{\gamma h}} = \frac{\partial J(W,b)}{\partial z_\gamma^t} a_h^{t-1}$$

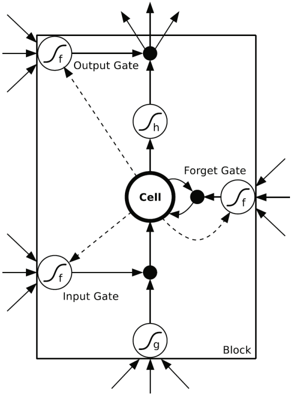
$$w_{\gamma c}: \frac{\partial J(W,b)}{\partial w_{\gamma c}} = \frac{\partial J(W,b)}{\partial z_\gamma^t}\frac{\partial z_\gamma^t}{\partial w_{\gamma c}} = \frac{\partial J(W,b)}{\partial z_\gamma^t} s_c^{t-1}$$

$$w_{hi}: \frac{\partial J(W,b)}{\partial w_{hi}} = \frac{\partial J(W,b)}{\partial z_h^t}\frac{\partial z_h^t}{\partial w_{hi}} = \frac{\partial J(W,b)}{\partial z_h^t} x_i^t$$

$$w_{hh}: \frac{\partial J(W,b)}{\partial w_{hh}} = \frac{\partial J(W,b)}{\partial z_h^t}\frac{\partial z_h^t}{\partial w_{hh}} = \frac{\partial J(W,b)}{\partial z_h^t} a_h^{t-1}$$

$$w_{kh}: \frac{\partial J(W,b)}{\partial w_{kh}} = \frac{\partial J(W,b)}{\partial z_k^t}\frac{\partial z_k^t}{\partial w_{kh}} = \frac{\partial J(W,b)}{\partial z_k^t} a_h^t$$

LSTM 单元



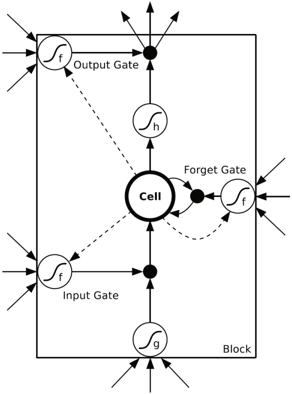
统计要计算的梯度：

$$\frac{\partial J(W,b)}{\partial z_\alpha^t} \quad \frac{\partial J(W,b)}{\partial z_\beta^t} \quad \frac{\partial J(W,b)}{\partial z_\gamma^t}$$

$$\frac{\partial J(W,b)}{\partial z_h^t} \quad \frac{\partial J(W,b)}{\partial z_k^t}$$

LSTM 单元

Step1：先解决 $\frac{\partial J(W,b)}{\partial z_k^t}$



若神经元 k 为输出层，则：

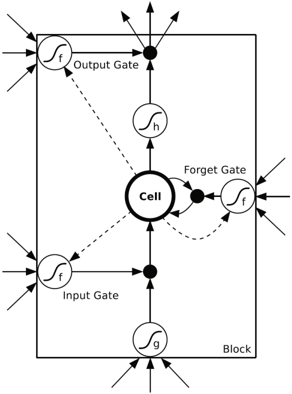
$$\frac{\partial J(W,b)}{\partial z_k^t} = \frac{\partial J(W,b)}{\partial a_k^t}\frac{\partial a_k^t}{\partial z_k^t} = \frac{\partial J(W,b)}{\partial a_k^t} output'(\cdot)$$

若神经元 k 的输出为下一时刻的输入，则：

$$\frac{\partial J(W,b)}{\partial z_k^t} = \frac{\partial J(W,b)}{\partial a_k^t}\frac{\partial a_k^t}{\partial z_k^t} = \sum_h^H \frac{\partial J(W,b)}{\partial z_h^{t+1}}\frac{\partial a_k^t}{\partial z_k^t} = \sum_h^H \frac{\partial J(W,b)}{\partial z_h^{t+1}} output'(\cdot)$$

LSTM 单元

Step 2：再解决 $\frac{\partial J(W,b)}{\partial z_\gamma^t}$



$$\frac{\partial J(W,b)}{\partial z_\gamma^t} = \sum_{h=1}^{H} \frac{\partial J(W,b)}{\partial a_h^t} \frac{\partial a_h^t}{\partial a_\gamma^t} \frac{\partial a_\gamma^t}{\partial z_\gamma^t}$$

因为：$a_h^t = a_\gamma^t h(s_c^t)$

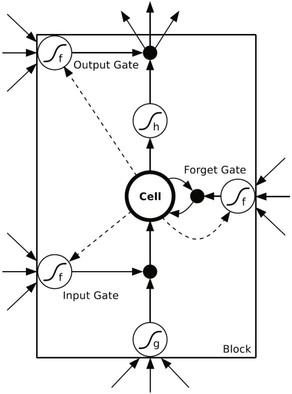
所以：

$$= \sum_{h=1}^{H} \frac{\partial J(W,b)}{\partial a_h^t} h(s_c^t) f'(z_\gamma^t)$$

$$= f'(z_\gamma^t) \sum_{h=1}^{H} \frac{\partial J(W,b)}{\partial a_h^t} h(s_c^t)$$

LSTM 单元

Step 2 继续：



$$\frac{\partial J(W,b)}{\partial a_h^t} = \sum_k^K \frac{\partial J(W,b)}{\partial z_k^t}\frac{\partial z_k^t}{\partial a_h^t} + \sum_h^H \frac{\partial J(W,b)}{\partial z_h^{t+1}}\frac{\partial z_h^{t+1}}{\partial a_h^t}$$

因为：

$$z_h^{t+1} = \sum_{i=1}^I w_{ci}x_i^{t+1} + \sum_{h=1}^H w_{hh}a_h^t$$

$$z_k^t = \sum_{k=1}^K w_{kh}a_h^t$$

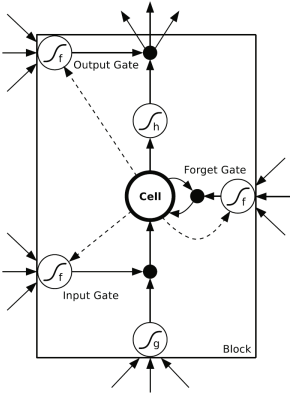
所以：

$$\frac{\partial J(W,b)}{\partial a_h^t} = \sum_k^K \frac{\partial J(W,b)}{\partial z_k^t}w_{kh} + \sum_h^H \frac{\partial J(W,b)}{\partial z_h^{t+1}}w_{hh}$$

其中：$\frac{\partial J(W,b)}{\partial z_k^t}$ 已解决，下面只需解决 $\frac{\partial J(W,b)}{\partial z_h^t}$

LSTM 单元

Step 3: 解决 $\frac{\partial J(W,b)}{\partial z_h^t}$



$$\frac{\partial J(W,b)}{\partial z_h^t} = \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} \frac{\partial s_c^t}{\partial z_h^t}$$

因为：$s_c^t = a_\alpha^t g(z_h^t) + a_\beta^t s_c^{t-1}$

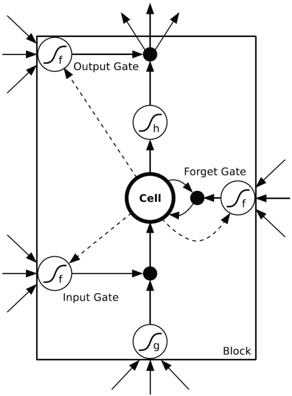
所以：

$$\frac{\partial J(W,b)}{\partial z_h^t} = \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} a_\alpha^t g'(z_h^t)$$

其中：$\frac{\partial J(W,b)}{\partial s_c^t}$ 待解决

LSTM 单元

Step 4: 解决 $\frac{\partial J(W,b)}{\partial z_\beta^t}$



$$\frac{\partial J(W,b)}{\partial z_\beta^t} = \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} \frac{\partial s_c^t}{\partial z_\beta^t}$$

因为：$s_c^t = a_\alpha^t g(z_h^t) + a_\beta^t s_c^{t-1}$

所以：

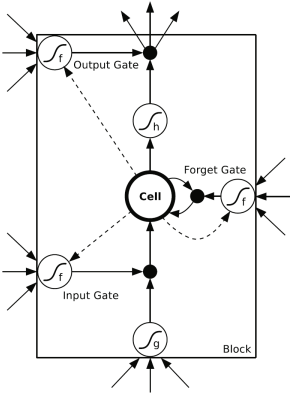
$$\frac{\partial J(W,b)}{\partial z_\beta^t} = \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} \frac{\partial a_\beta^t}{\partial z_\beta^t} s_c^{t-1}$$

$$= f'(z_\beta^t) \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} s_c^{t-1}$$

其中：$\frac{\partial J(W,b)}{\partial s_c^t}$ 待解决

LSTM 单元

Step 5: 解决 $\frac{\partial J(W,b)}{\partial z_\alpha^t}$



$$\frac{\partial J(W,b)}{\partial z_\alpha^t} = \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} \frac{\partial s_c^t}{\partial z_\alpha^t}$$

因为：$s_c^t = a_\alpha^t g(z_h^t) + a_\beta^t s_c^{t-1}$

所以：

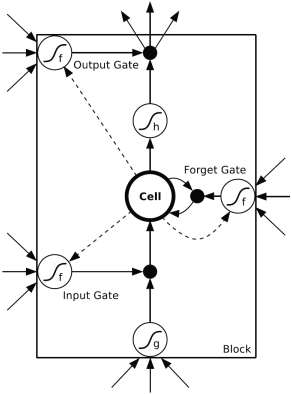
$$\frac{\partial J(W,b)}{\partial z_\alpha^t} = \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} \frac{\partial a_\alpha^t}{\partial z_\alpha^t} g(z_h^t)$$

$$= f'(z_\alpha^t) \sum_{c=1}^{C} \frac{\partial J(W,b)}{\partial s_c^t} g(z_h^t)$$

其中：$\frac{\partial J(W,b)}{\partial s_c^t}$ 待解决

LSTM 单元

Step 6: 解决焦点 $\frac{\partial J(W,b)}{\partial s_c^t}$

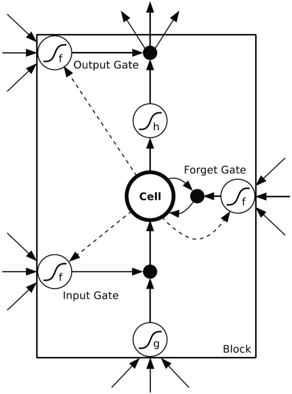


因为：

$$\frac{\partial J(W,b)}{\partial s_c^t} = \frac{\partial J(W,b)}{\partial a_c^t}\frac{\partial a_c^t}{\partial s_c^t} + \frac{\partial J(W,b)}{\partial s_c^{t+1}}\frac{\partial s_c^{t+1}}{\partial s_c^t} + \frac{\partial J(W,b)}{\partial z_\alpha^{t+1}}\frac{\partial z_\alpha^{t+1}}{\partial s_c^t} + \frac{\partial J(W,b)}{\partial z_\beta^{t+1}}\frac{\partial z_\beta^{t+1}}{\partial s_c^t} + \frac{\partial J(W,b)}{\partial z_\gamma^{t+1}}\frac{\partial z_\gamma^{t+1}}{\partial s_c^t}$$

LSTM 单元

Step 6: 继续焦点 $\frac{\partial J(W,b)}{\partial s_c^t}$



又因为：$a_h^t = a_\gamma^t h(s_c^t) = a_\gamma^t a_c^t$

故：
$$\frac{\partial J(W,b)}{\partial a_c^t}\frac{\partial a_c^t}{\partial s_c^t} = \frac{\partial J(W,b)}{\partial a_h^t}\frac{\partial a_h^t}{\partial s_c^t} = \frac{\partial J(W,b)}{\partial a_h^t} a_\gamma^t h'(s_c^t)$$

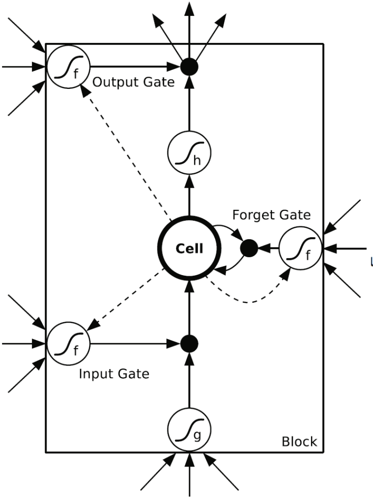
所以：
$$\frac{\partial J(W,b)}{\partial s_c^t} = \frac{\partial J(W,b)}{\partial a_h^t} a_\gamma^t h'(s_c^t) + \frac{\partial J(W,b)}{\partial s_c^{t+1}} a_\beta^{t+1} + \frac{\partial J(W,b)}{\partial z_\alpha^{t+1}} w_{\alpha c} + \frac{\partial J(W,b)}{\partial z_\beta^{t+1}} w_{\beta c} + \frac{\partial J(W,b)}{\partial z_\gamma^{t+1}} w_{\gamma c}$$

LSTM with a forget gate [edit]

Compact form of the equations for the forward pass of an LSTM unit with a fo

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

where the initial values are $c_0 = 0$ and $h_0 = 0$ and the operator $\circ$ denotes t



Variables [edit]

- $x_t \in R^d$: input vector to the LSTM unit
- $f_t \in R^h$: forget gate's activation vector
- $i_t \in R^h$: input gate's activation vector
- $o_t \in R^h$: output gate's activation vector
- $h_t \in R^h$: output vector of the LSTM unit
- $c_t \in R^h$: cell state vector
- $W \in R^{h \times d}$, $U \in R^{h \times h}$ and $b \in R^h$: weight matrices and bias vector parameters which need to be learned during training

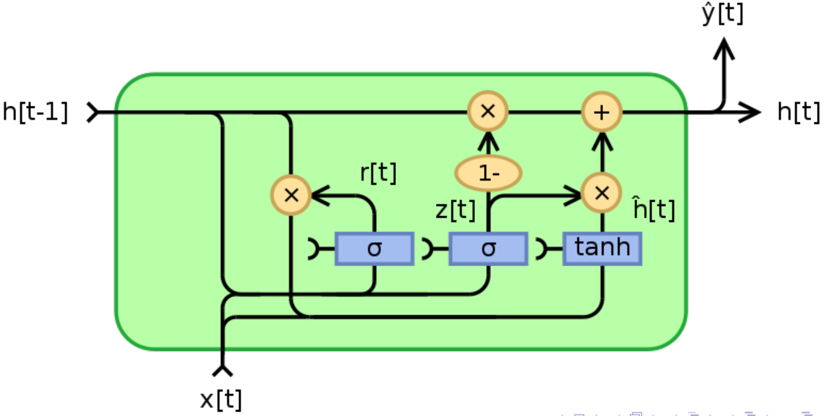
where the superscripts d and h refer to the number of input features and number of hidden units, respectively.

Activation functions [edit]

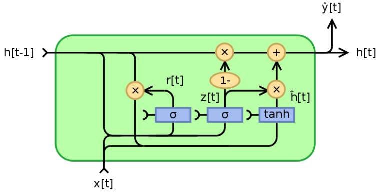
- σ_g: sigmoid function.
- σ_c: hyperbolic tangent function.
- σ_h: hyperbolic tangent function or, as the peephole LSTM paper[which?] suggests, $\sigma_h(x) = x$.[17][18]

GRU

Gated recurrent units (GRUs) are a gating mechanism in recurrent neural networks, introduced in 2014. Their performance was found to be similar to that of long short-term memory (LSTM). However, GRUs have been shown to exhibit better performance on smaller datasets.



GRU



Initially, for $t = 0$, the output vector is $h_0 = 0$.

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r)$$
$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h)$$

Variables

- x_t: input vector
- h_t: output vector
- z_t: update gate vector
- r_t: reset gate vector
- W, U and b: parameter matrices and vector

Activation functions

- σ_g: The original is a sigmoid function.
- σ_h: The original is a hyperbolic tangent.

Thanks.