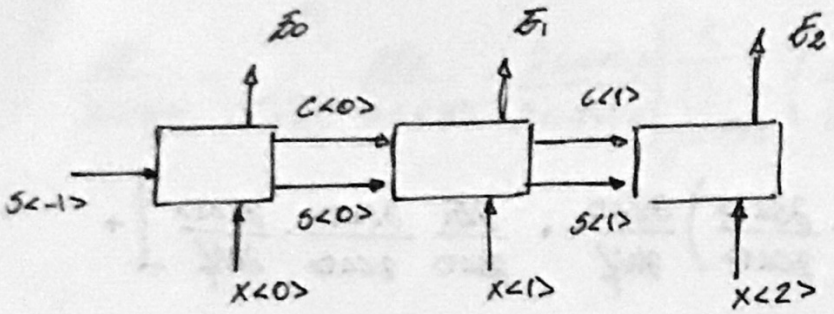
LSTM Network gradient through time



$$\frac{\partial E}{\partial W_f} = \left[\frac{\partial E_2}{\partial S_{<2>}} \cdot \frac{\partial S_{<2>}}{\partial C_{<2>}} \cdot \frac{\partial C_{<2>}}{\partial W_f}\right] + \left[\frac{\partial E_2}{\partial S_{<2>}} \cdot \frac{\partial S_{<2>}}{\partial C_{<2>}} \cdot \frac{\partial C_{<2>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial W_{ax}} +\right.$$

$$+ \frac{\partial E_2}{\partial S_{<2>}} \cdot \frac{\partial S_{<2>}}{\partial S_{<2>}} \cdot \frac{\partial C_{<2>}}{\partial S_{<1>}} \cdot \frac{\partial S_{<1>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial W_f} + \frac{\partial E_1}{\partial S_{<1>}} \cdot \frac{\partial S_{<1>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial W_f}\Big] +$$

$$+ \left[\frac{\partial E_2}{\partial S_{<2>}} \cdot \frac{\partial S_{<2>}}{\partial C_{<2>}} \cdot \frac{\partial C_{<2>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial C_{<0>}} \cdot \frac{\partial C_{<0>}}{\partial W_f} +\right.$$

$$+ \frac{\partial E_2}{\partial S_{<2>}} \cdot \frac{\partial S_{<2>}}{\partial C_{<2>}} \cdot \frac{\partial C_{<2>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial S_{<0>}} \cdot \frac{\partial S_{<0>}}{\partial C_{<0>}} \cdot \frac{\partial C_{<0>}}{\partial W_f}$$

$\partial S_{<1>}/\partial C_{<1>}$

$$+ \frac{\partial E_2}{\partial S_{<2>}} \cdot \frac{\partial S_{<2>}}{\partial C_{<2>}} \cdot \frac{\partial C_{<2>}}{\partial S_{<1>}} \cdot \frac{\partial S_{<1>}}{\partial S_{<0>}} \cdot \frac{\partial S_{<0>}}{\partial W_f} +$$

$$+ \frac{\partial E_2}{\partial S_{<2>}} \cdot \frac{\partial S_{<2>}}{\partial C_{<2>}} \cdot \frac{\partial C_{<2>}}{\partial S_{<1>}} \cdot \frac{\partial S_{<1>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial S_{<0>}} \cdot \frac{\partial S_{<0>}}{\partial C_{<0>}} \cdot \frac{\partial C_{<0>}}{\partial W_f}$$

$$+ \frac{\partial E_1}{\partial S_{<1>}} \cdot \frac{\partial S_{<1>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial C_{<0>}} \cdot \frac{\partial C_{<0>}}{\partial W_f} + \frac{\partial E_1}{\partial S_{<1>}} \cdot \frac{\partial S_{<1>}}{\partial C_{<1>}} \cdot \frac{\partial C_{<1>}}{\partial C_{<0>}} \cdot \frac{\partial C_{<0>}}{\partial W_f}$$

$$+ \frac{\partial E_0}{\partial S_{<0>}} \cdot \frac{\partial S_{<0>}}{\partial C_{<0>}} \cdot \frac{\partial C_{<0>}}{\partial W_f}$$

$$= \frac{\partial E_2}{\partial s^{<2>}} \cdot \frac{\partial s^{<2>}}{\partial c^{<2>}} \cdot \frac{\partial c^{<2>}}{\partial w_f} +$$

$$+ \left[\frac{\partial E_2}{\partial s^{<2>}} \cdot \frac{\partial s^{<2>}}{\partial c^{<2>}} \left(\frac{\partial c^{<2>}}{\partial c^{<1>}} + \frac{\partial c^{<2>}}{\partial s^{<1>}} \cdot \frac{\partial s^{<1>}}{\partial c^{<1>}} \right) \frac{\partial c^{<1>}}{\partial w_f} + \frac{\partial E_1}{\partial s^{<1>}} \frac{\partial s^{<1>}}{\partial c^{<1>}} \cdot \frac{\partial c^{<1>}}{\partial w_f} \right] +$$

$$+ \left[\frac{\partial E_2}{\partial s^{<2>}} \cdot \frac{\partial s^{<2>}}{\partial c^{<2>}} \left(\frac{\partial c^{<2>}}{\partial c^{<1>}} + \frac{\partial c^{<2>}}{\partial s^{<1>}} \cdot \frac{\partial s^{<1>}}{\partial c^{<1>}} \right) \left(\frac{\partial c^{<1>}}{\partial c^{<0>}} + \frac{\partial c^{<1>}}{\partial s^{<0>}} \cdot \frac{\partial s^{<0>}}{\partial c^{<0>}} \right) \frac{\partial c^{<0>}}{\partial w_f} \right.$$

$$+ \frac{\partial E_1}{\partial s^{<1>}} \cdot \frac{\partial s^{<1>}}{\partial c^{<1>}} \left(\frac{\partial c^{<1>}}{\partial c^{<0>}} + \frac{\partial c^{<1>}}{\partial s^{<0>}} \cdot \frac{\partial s^{<0>}}{\partial c^{<0>}} \right) \frac{\partial c^{<0>}}{\partial w_f} +$$

$$\left. + \frac{\partial E_1}{\partial s^{<0>}} \cdot \frac{\partial s^{<0>}}{\partial c^{<0>}} \cdot \frac{\partial c^{<0>}}{\partial w_f} \right] \Rightarrow$$

$$\frac{\partial E}{\partial w_f} = \sum_{l=0}^{T_x} \sum_{k=l}^{T_y} \frac{\partial E_k}{\partial s^{<k>}} \cdot \frac{\partial s^{<k>}}{\partial c^{<k>}} \left[\prod_{j=l+1}^{k} \left(\frac{\partial c^{<j>}}{\partial c^{<j-1>}} + \frac{\partial c^{<j>}}{\partial s^{<j-1>}} \cdot \frac{\partial s^{<j-1>}}{\partial c^{<j-1>}} \right) \right] \frac{\partial c^{<l>}}{\partial w_f}$$

$$\frac{\partial E}{\partial w_a} = \sum_{l=0}^{T_x} \sum_{k=l}^{T_y} \frac{\partial E_k}{\partial s^{<k>}} \cdot \frac{\partial s^{<k>}}{\partial c^{<k>}} \left[\prod_{j=l+1}^{k} \left(\frac{\partial c^{<j>}}{\partial c^{<j-1>}} + \frac{\partial c^{<j>}}{\partial s^{<j-1>}} \cdot \frac{\partial s^{<j-1>}}{\partial c^{<j-1>}} \right) \right] \frac{\partial c^{<l>}}{\partial w_a}$$

$$\frac{\partial E}{\partial w_c} = \sum_{l=0}^{T_x} \sum_{k=l}^{T_y} \frac{\partial E_k}{\partial s^{<k>}} \cdot \frac{\partial s^{<k>}}{\partial c^{<k>}} \left[\prod_{j=l+1}^{k} \left(\frac{\partial c^{<j>}}{\partial c^{<j-1>}} + \frac{\partial c^{<j>}}{\partial s^{<j-1>}} \cdot \frac{\partial s^{<j-1>}}{\partial c^{<j-1>}} \right) \right] \frac{\partial c^{<l>}}{\partial w_c}$$

$$\frac{\partial E}{\partial w_b} = \sum_{l=0}^{T_x} \sum_{k=l}^{T_y} \frac{\partial E_k}{\partial s^{<k>}} \cdot \prod_{j=l+1}^{k} \left(\frac{\partial s^{<j>}}{\partial s^{<j-1>}} \right) \frac{\partial s^{<l>}}{\partial w_b}$$

$$\frac{\partial c^{<j>}}{\partial c^{<j-1>}} = \Gamma_f^{<t>} \rightarrow \prod \frac{\partial c^{<j>}}{\partial c^{<j-1>}}$$

will not vanish or explode if the Γ_f is on the linear part $\rightarrow$ Value $(0,1)$ depending if it had importance $\rightarrow$ If enabled and Γ_f enabled update.

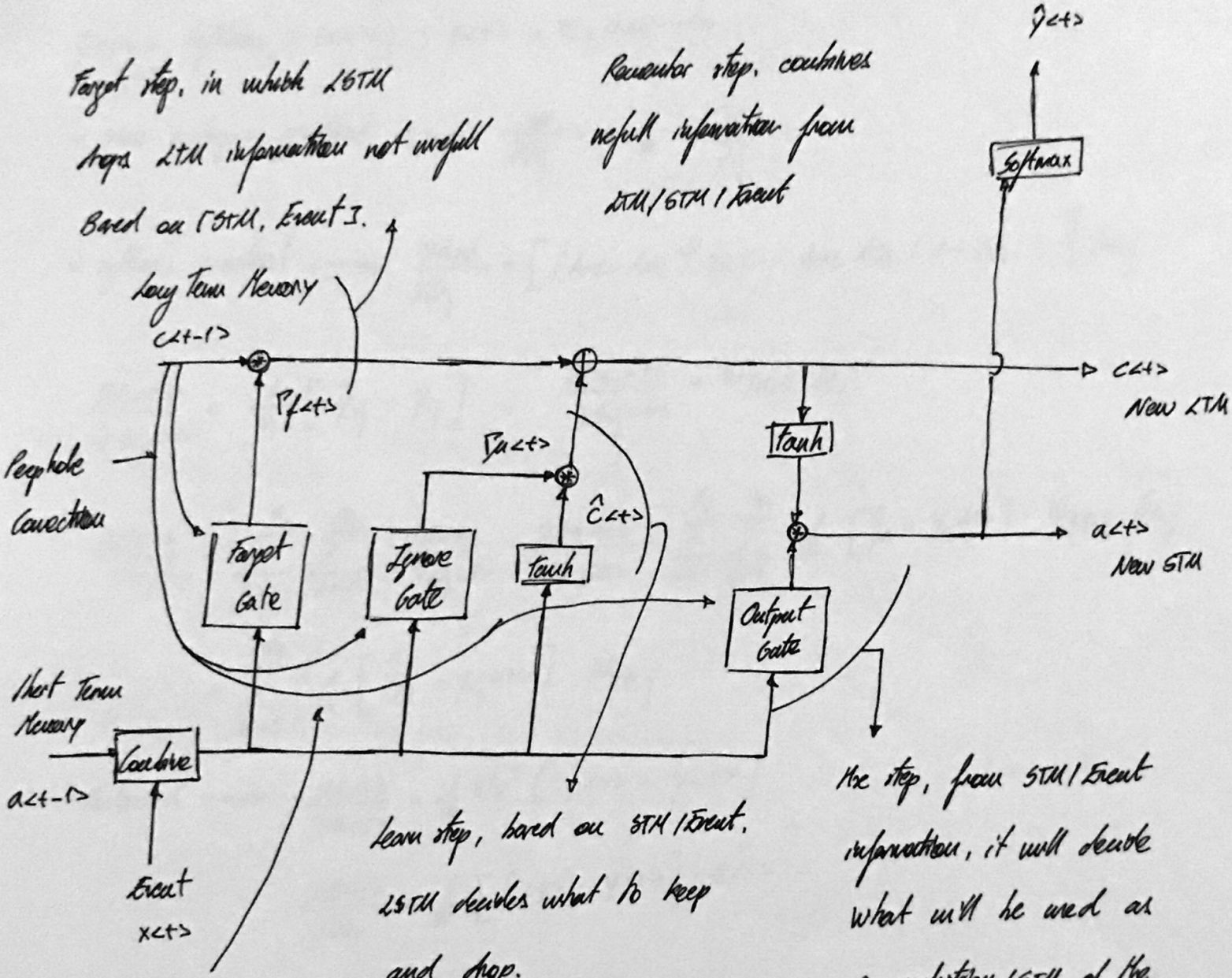
$\rightarrow$ for this gradient, it won't vanish because of the path until previous block $\frac{\partial c^{<t+1>}}{\partial a^{<t>}} \cdot \frac{\partial a^{<t>}}{\partial w_b}$

$$\frac{\partial E}{\partial x_{\langle l \rangle}} = \sum_{k=l}^{T_y} \frac{\partial E_k}{\partial s_{\langle k \rangle}} \cdot \frac{\partial s_{\langle k \rangle}}{\partial c_{\langle k \rangle}} \left[\prod_{j=l+1}^{k} \left(\frac{\partial c_{\langle j \rangle}}{\partial c_{\langle j-1 \rangle}} + \frac{\partial c_{\langle j \rangle}}{\partial s_{\langle j-1 \rangle}} \cdot \frac{\partial s_{\langle j-1 \rangle}}{\partial c_{\langle j-1 \rangle}} \right) \right] \frac{\partial c_{\langle l \rangle}}{\partial x_{\langle l \rangle}}$$

Long Short Term Memory

Forget step, in which LSTM keeps LTM information not useful. Based on [STM, Event].

Remember step, combines useful information from LTM / STM / Event.



Learn step, based on STM / Event. LSTM decides what to keep and drop. Useful information of STM / Event.

Use step, from STM / Event information, it will decide what will be used as a prediction / STM of the combination LTM / STM / Event.

Peephole connections: LSTM variant, includes LTM in gate decisions.

$$\Gamma f\langle t\rangle = \sigma(Wf[a\langle t-1\rangle, x\langle t\rangle] + bf)$$

$$\Gamma u\langle t\rangle = \sigma(Wu[a\langle t-1\rangle, x\langle t\rangle] + bu)$$

$$\hat{c}\langle t\rangle = \tanh(Wc[a\langle t-1\rangle, x\langle t\rangle] + bc)$$

$$c\langle t\rangle = c\langle t-1\rangle * \Gamma f\langle t\rangle + \Gamma u\langle t\rangle * \hat{c}\langle t\rangle$$

$$\Gamma o\langle t\rangle = \sigma(Wo[a\langle t-1\rangle, x\langle t\rangle] + bo)$$

$$a\langle t\rangle = \Gamma o\langle t\rangle * \tanh(c\langle t\rangle)$$

cell backprop →

$\hat{y}^{\langle t \rangle} = \text{softmax}(z^{\langle t \rangle})$; $z^{\langle t \rangle} = W_y a^{\langle t \rangle} + b_y$

* cross-entropy gradient → $\frac{\partial E}{\partial \hat{y}_{ij}} = -\frac{1}{m} \cdot \frac{y_{ij}}{\hat{y}_{ij}}$

* softmax gradient → $\frac{\partial A_{re}}{\partial z_{ij}} = \left[(A_{re} - A_{re})^2 \delta_{ki} - A_{re} A_{ie} (1 - \delta_{ki}) \right] \delta_{ej}$

$\frac{\partial E^{\langle t \rangle}}{\partial z_{ij}^{\langle t \rangle}} = \frac{1}{m} \left[\hat{y}_{ij} - y_{ij} \right]$; $\frac{\partial z_{re}^{\langle t \rangle}}{\partial a_{ij}^{\langle t \rangle}} = W_{yk,i}\, \delta_{ej}$

$$\frac{\partial E^{\langle t \rangle}}{\partial a_{ij}^{\langle t \rangle}} = \sum_{k=1}^{n_y} \sum_{s=1}^{m} \frac{\partial E^{\langle t \rangle}}{\partial z_{kj}^{\langle t \rangle}} \cdot \frac{\partial z_{rs}^{\langle t \rangle}}{\partial a_{ij}^{\langle t \rangle}} = \sum_{k=1}^{n_y} \sum_{s=1}^{m} \frac{1}{m} \left[\hat{y}_{rs} - y_{rs}^{\langle t \rangle} \right] \cdot W_{yr,i}\, \delta_{s,j}$$

$$= \sum_{k=1}^{n_y} \frac{1}{m} \left[\hat{y}_{rj} - y_{rj}^{\langle t \rangle} \right] \cdot W_{yr,i}$$

Vectorized → $\frac{\partial E^{\langle t \rangle}}{\partial a^{\langle t \rangle}} = \frac{1}{m} W_y^T \left[\hat{y}^{\langle t \rangle} - y^{\langle t \rangle} \right]$

$$\frac{\partial E^{\langle t \rangle}}{\partial W_y} = \frac{1}{m} \left[\hat{y}^{\langle t \rangle} - y^{\langle t \rangle} \right] \cdot a^{T\langle t \rangle}$$

$\frac{\partial C_{re}^{\langle t \rangle}}{\partial C_{ij}^{\langle t-1 \rangle}} = (\not{4} + \Gamma_{A_{re}}^{\langle t \rangle})\, \delta_{ki}\, \delta_{ej}$; $\frac{\partial a_{re}^{\langle t \rangle}}{\partial C_{ij}^{\langle t \rangle}} = \Gamma_{o_{re}}^{\langle t \rangle} \left(1 - \tanh^2(C_{re}^{\langle t \rangle})\right) \delta_{ki}\, \delta_{ej}$

$$\frac{\partial E^{\text{recurrent}}}{\partial C_{ij}^{\langle t \rangle}} = \frac{\partial E^{\langle t+1:T_y \rangle}}{\partial C_{ij}^{\langle t \rangle}} + \sum_{r=1}^{n_c} \sum_{s=1}^{m} \left(\frac{\partial E^{\langle t+1:T_y \rangle}}{\partial a_{rs}^{\langle t \rangle}} + \frac{\partial E^{\langle t \rangle}}{\partial a_{rs}^{\langle t \rangle}} \right) \Gamma_{o_{rs}}^{\langle t \rangle} \left(1 - \tanh^2(C_{rs}^{\langle t \rangle})\right) \delta_{ki}\, \delta_{sj}$$

$$= \frac{\partial E^{\langle t+1:T_y \rangle}}{\partial C_{ij}^{\langle t \rangle}} + \left[\frac{\partial E^{\langle t+1:T_y \rangle}}{\partial a_{ij}^{\langle t \rangle}} + \frac{\partial E^{\langle t \rangle}}{\partial a_{ij}^{\langle t \rangle}} \right] \Gamma_{o_{ij}}^{\langle t \rangle} \left(1 - \tanh^2(C_{ij}^{\langle t \rangle})\right)$$

Vectorized → $\frac{\partial E}{\partial C^{\langle t+m \rangle}} = \frac{\partial E^{\langle t+1:T_y \rangle}}{\partial C^{\langle t \rangle}} + \left[\frac{\partial E^{\langle t+1:T_y \rangle}}{\partial a^{\langle t \rangle}} + \frac{\partial E^{\langle t \rangle}}{\partial a^{\langle t \rangle}} \right] * \Gamma_o^{\langle t \rangle} * \left(1 - \tanh^2(C^{\langle t \rangle})\right)$

$$\frac{\partial E}{\partial C_{ij}^{<t-1>}} = \sum_{k=1}^{n_c} \sum_{l=1}^{m} \frac{\partial E}{\partial C_{kl}^{<t>}} \cdot \underbrace{\Gamma_{f_{kl}}^{<t>} \delta_{ki} \delta_{lj}}_{\partial C_{kl}^{<t>} / \partial C_{ij}^{<t-1>}} = \frac{\partial E}{\partial C_{ij}^{<t>}} \cdot \Gamma_{f_{ij}}^{<t>}$$

Vectorized $\longrightarrow$ $\frac{\partial E}{\partial C^{<t-1>}} = \frac{\partial E}{\partial C^{<t>}} * \Gamma_f^{<t>}$

$$\frac{\partial E}{\partial \Gamma_{f_{ij}}^{<t>}} = \frac{\partial E}{\partial C_{ij}^{<t>}} \cdot C_{ij}^{<t-1>} \; ; \quad \frac{\partial E}{\partial \Gamma_{u}(f)} = \frac{\partial E}{\partial C_{ij}^{<t>}} * \hat{C}_{ij}^{<t>} \; ; \quad \frac{\partial E}{\partial \hat{C}_{ij}^{<t>}} = \frac{\partial E}{\partial C_{ij}^{<t>}} * \Gamma_{u_{ij}}^{<t>}$$

Vectorized $\longrightarrow$ $\frac{\partial E}{\partial \Gamma_f^{<t>}} = \frac{\partial E}{\partial C^{<t>}} * C^{<t-1>} \; ; \; \frac{\partial E}{\partial \Gamma_u^{<t>}} = \frac{\partial E}{\partial C^{<t>}} * \hat{C}^{<t>} \; ; \; \frac{\partial E}{\partial \hat{C}^{<t>}} = \frac{\partial E}{\partial C^{<t>}} * \Gamma_u^{<t>}$

$$\frac{\partial E}{\partial \Gamma_{o_{ij}}^{<t>}} = \sum_{r=1}^{n_c} \sum_{s=1}^{m} \left[\frac{\partial E^{<t>}}{\partial a_{rs}^{<t>}} + \frac{\partial E^{<t+1:T_y>}}{\partial a_{rs}^{<t>}} \right] \cdot \tanh\left(C_{ij}^{<t>}\right) \delta_{ri} \delta_{sj}$$

Vectorized $\longrightarrow$ $\frac{\partial E}{\partial \Gamma_o^{<t>}} = \frac{\partial E}{\partial a^{<t>}} * \tanh\left(C^{<t>}\right) \; ; \; \frac{\partial E}{\partial W_o} = \left[\frac{\partial E}{\partial a^{<t>}} * \tanh\left(C^{<t>}\right) \right] \left[a^{<t-1>}, x^{<t>} \right]^T$

$$\frac{\partial E}{\partial \left[a^{<t-1>}, x^{<t>} \right]} = \sum_{r=1}^{n_c} \sum_{s=1}^{m} \left[\frac{\partial E}{\partial \Gamma_{f_{rs}}^{<t>}} \cdot W_{f_{ri}} \delta_{sj} + \frac{\partial E}{\partial \Gamma_{u_{rs}}^{<t>}} \cdot W_{u_{ri}} \delta_{sj} + \frac{\partial E}{\partial \hat{C}^{<t>}} W_{c_{ri}} \delta_{sj} \right.$$

$$\left. + \frac{\partial E}{\partial \Gamma_{o_{rj}}^{<t>}} \cdot W_{o_{ri}} \cdot \delta_{sj} \right] = \sum_{r=1}^{n_c} \left[\frac{\partial E}{\partial \Gamma_{f_{rj}}^{<t>}} \cdot W_{f_{ri}} + \frac{\partial E}{\partial \Gamma_{u_{rj}}} \cdot W_{u_{ri}} + \right.$$

$$\left. + \frac{\partial E}{\partial \Gamma_{o_{rj}}^{<t>}} W_{o_{ri}} + \frac{\partial E}{\partial \hat{C}_{rj}^{<t>}} W_{c_{ri}} \right]$$

Vectorized

$$\frac{\partial E}{\partial \left[a^{<t-1>}, x^{<t>} \right]} = \frac{\partial E}{\partial \Gamma_f^{<t>}} \cdot W_f + W_u^T \cdot \frac{\partial E}{\partial \Gamma_u^{<t>}} + W_c^T \frac{\partial E}{\partial \hat{C}^{<t>}} + W_o^T \left(\frac{\partial E}{\partial \Gamma_o^{<t>}} * \Gamma_o^{<t>} * \left(1 - \Gamma_o^{<t>}\right) \right.$$

$\left.\right)$ ← $* \Gamma_f (1 - \Gamma_f)$ ← $* \Gamma_u (1 - \Gamma_u)$ ← $(1 - \tanh^2(\cdot))$ $(1 - \hat{C}^{<t>})$ ← $(\cdots)$

sigmoids

$\left[a^{<t-1>}, x^{<t>} \right] \equiv \text{data}\left(n_a + n_x, m\right)$

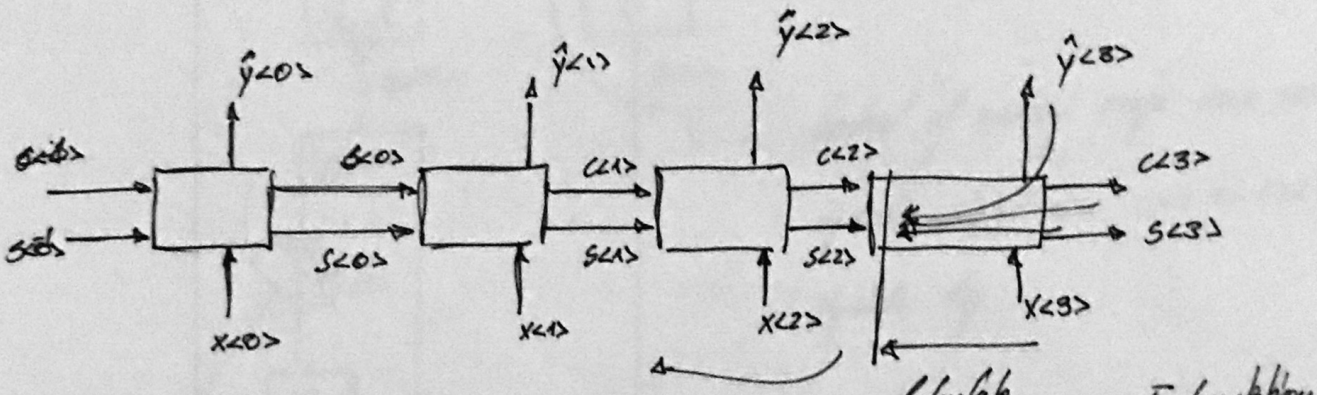
$$\longmapsto \frac{\partial E}{\partial a^{<t-1>}} = \frac{\partial E}{\partial \left[a^{<t-1>}, x^{<t>} \right]} \left[: n_a, m \right] \; ; \; \frac{\partial E}{\partial x^{<t>}} = \frac{\partial E}{\left[\ldots \right]} \left[n_a : n_a + n_x, m \right]$$

Rückwärts ⟶

$$\frac{\partial E}{\partial W_f} = \left[\frac{\partial E}{\partial \Gamma_f^{\langle t \rangle}} * \Gamma_f^{\langle t \rangle} * (1 - \Gamma_f^{\langle t \rangle}) \right] \cdot \left[a^{\langle t-1 \rangle}, x^{\langle t \rangle} \right]^T$$

$$\frac{\partial E}{\partial W_u} = \left[\frac{\partial E}{\partial \Gamma_u^{\langle t \rangle}} * \Gamma_u^{\langle t \rangle} * (1 - \Gamma_u^{\langle t \rangle}) \right] \cdot \left[a^{\langle t-1 \rangle}, x^{\langle t \rangle} \right]^T$$

$$\frac{\partial E}{\partial W_c} = \left[\frac{\partial E}{\partial \tilde{c}^{\langle t \rangle}} * (1 - \hat{c}^{2\langle t \rangle}) \right] \cdot \left[a^{\langle t-1 \rangle}, x^{\langle t \rangle} \right]^T$$



calculate grad; Time step une.

Implementation