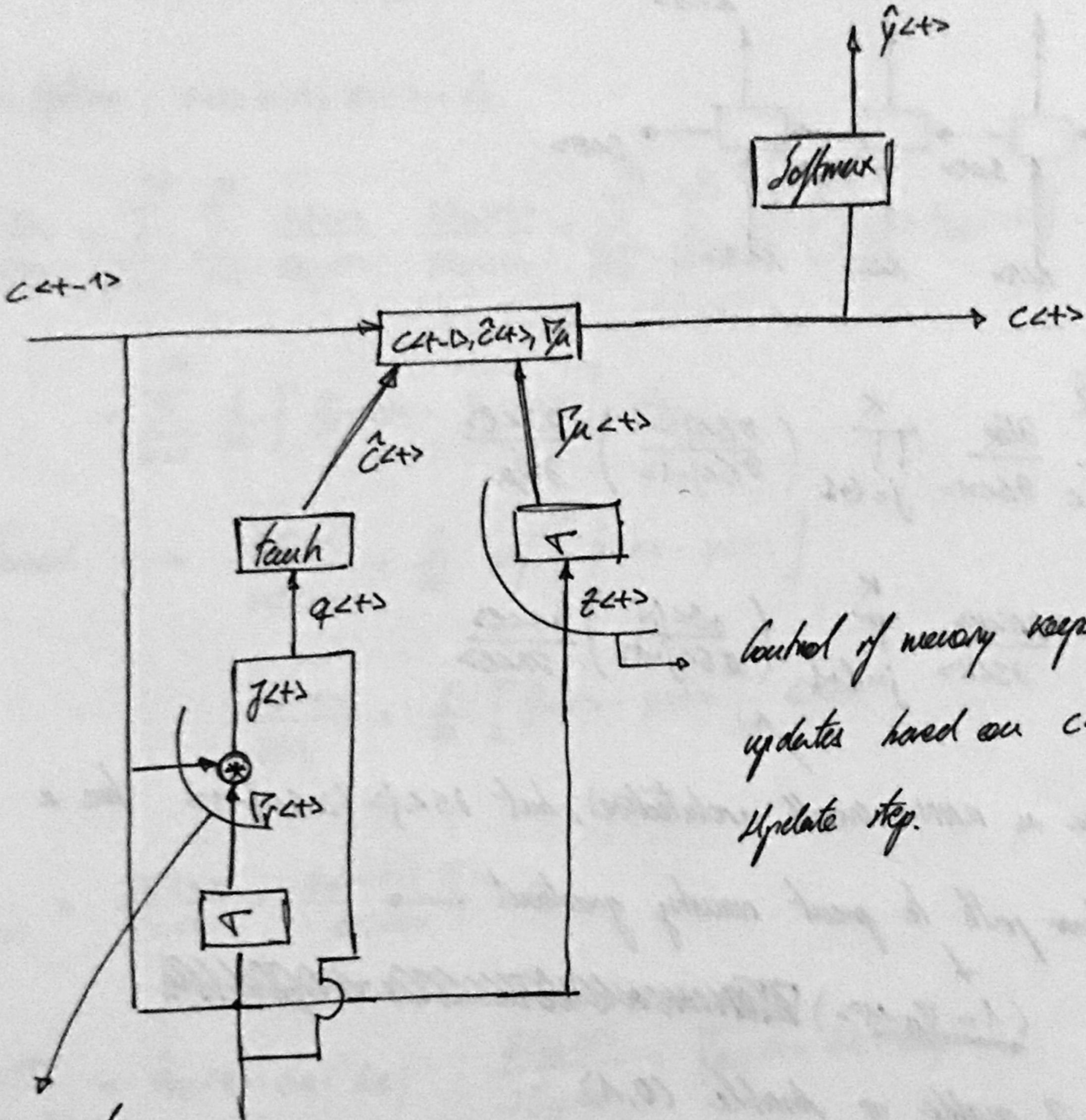
Gated Recurrent Unit



Control of memory keeps same values or updates based on c<t-1> / x<t>.

Update step.

Reset, relevance step.

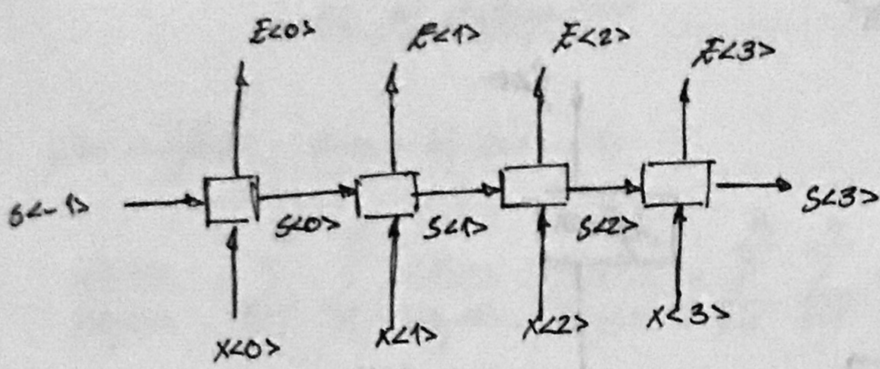
Decides which features ~~to~~ of memory we want to keep.

$$\Gamma_r^{<t>} = \sigma\left(W_r\left[c^{<t-1>}, x^{<t>}\right] + b_r\right)$$

$$\hat{c}^{<t>} = \tanh\left(W_c\left[\Gamma_r^{<t>} * c^{<t-1>}, x^{<t>}\right] + b_c\right)$$

$$\Gamma_u^{<t>} = \sigma\left(W_u\left[c^{<t-1>}, x^{<t>}\right] + b_u\right)$$

$$c^{<t>} = \Gamma_u^{<t>} * \hat{c}^{<t>} + (1 - \Gamma_u^{<t>}) * c^{<t-1>}$$



$$\frac{\partial E}{\partial W_u} = \sum_{l=0}^{T_x} \sum_{k=l}^{T_y} \frac{\partial E_k}{\partial S\langle k\rangle} \prod_{j=l+1}^{k} \left(\frac{\partial S\langle j\rangle}{\partial S\langle j-1\rangle} \right) \frac{\partial S\langle l\rangle}{\partial W_u}$$

$$\frac{\partial E}{\partial X\langle l\rangle} = \sum_{k=l}^{T_y} \frac{\partial E\langle k\rangle}{\partial S\langle k\rangle} \prod_{j=l+1}^{k} \left(\frac{\partial S\langle j\rangle}{\partial S\langle j-1\rangle} \right) \frac{\partial S\langle l\rangle}{\partial X\langle l\rangle}$$

Same as RNN (overall architecture), but $\partial S\langle j\rangle / \partial S\langle j-1\rangle$ has a linear path to prevent vanishing gradient $\longrightarrow$

$(1 - \Gamma_u\langle t\rangle)$ ~~[illegible]~~

Γ enable or disable $(0,1)$

If enabled and $E\uparrow$, path to carry update. If disable and $E\uparrow$,

W_u will change to minimize $\longrightarrow$ carry gradient when enabled.

GRU cell Backprop $\longrightarrow$

$\hat{y}^{<t>} = \text{softmax}$; $z^{<t>} = W_y a^{<t>} + b_y$

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

$$= \sum_{k=1}^{n_y} \frac{1}{m} \left[\hat{y}_{rj}^{<t>} - y_{rj}^{<t>} \right] \cdot W_{y_{r,j}}$$

Vectorized $\longrightarrow$ $\dfrac{\partial E^{<t>}}{\partial a^{<t>}} = \dfrac{1}{m} W_y^T \left[\hat{y}^{<t>} - y^{<t>} \right]$

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

$$\frac{\partial E}{\partial c^{<t>}} = \frac{\partial E^{<t>}}{\partial c^{<t>}} + \frac{\partial E^{<t+1:T_y>}}{\partial c^{<t>}}$$

$$\frac{\partial c_{kl}^{<t>}}{\partial \Gamma_{u,ij}^{<t>}} = \hat{c}_{kl}^{<t>} \delta_{k,i} \delta_{l,j} \quad ; \quad \frac{\partial c_{kl}^{<t>}}{\partial \hat{c}_{ij}^{<t>}} = \Gamma_{u,ij}^{<t>} \delta_{k,i} \delta_{l,j}$$

$$\frac{\partial E}{\partial \Gamma_u^{<t>}} = \frac{\partial E}{\partial c^{<t>}} * \hat{c}^{<t>} \quad ; \quad \frac{\partial E}{\partial \hat{c}^{<t>}} = \frac{\partial E}{\partial c^{<t>}} * \Gamma_u^{<t>} \quad \longleftarrow \text{Vectorized.}$$

$$\frac{\partial E}{\partial t_{ij}^{<t>}} = \sum_{r=1}^{n_c} \sum_{s=1}^{m} \left(\frac{\partial E}{\partial \Gamma_u^{<t>}} * \Gamma_{u_{rs}}^{<t>} \cdot (1 - \Gamma_{u_{rs}}^{<t>}) \right) \cdot W_{u_{r,i}} \delta_{s,j}$$

$$\frac{\partial E}{\partial t^{<t>}} = W_u^T \left[\frac{\partial E}{\partial \Gamma^{<t>}} * \Gamma_u^{<t>} * (1 - \Gamma_u^{<t>}) \right]$$

$\longleftarrow$ Vectorized.

$$\frac{\partial E}{\partial q^{<t>}} = W_c^T \left[\frac{\partial E}{\partial \hat{c}^{<t>}} * (1 - \hat{c}^{2<t>}) \right]$$

$$J^{\langle t \rangle} = q^{\langle t \rangle}[:n_c, :] \longrightarrow \text{Just elements corresponding to } c^{\langle t-1 \rangle}$$

$$\frac{\partial E}{J^{\langle t \rangle}} = \frac{\partial E}{\partial q^{\langle t \rangle}}[:n_c, :]$$

$$J^{\langle t \rangle} = \Gamma_r^{\langle t \rangle} * c^{\langle t-1 \rangle} \longrightarrow \frac{\partial J^{\langle t \rangle}_{kl}}{\partial \Gamma_{r_{ij}}^{\langle t \rangle}} = c_{kl}^{\langle t-1 \rangle}\, \delta_{ki}\, \delta_{jl} \; ; \; \frac{\partial J_{kl}^{\langle t \rangle}}{\partial c_{ij}^{\langle t-1 \rangle}} = \Gamma_{r_{kl}}^{\langle t-1 \rangle}\, \delta_{ki} \delta_{lj}$$

$$\frac{\partial E}{\partial p^{\langle t \rangle}} = \cancel{(\partial E)}\; W_r^T \left[\frac{\partial E}{\partial J^{\langle t \rangle}} * c_k^{\langle t \rangle} * \Gamma_r^{\langle t \rangle} * (1 - \Gamma_r^{\langle t \rangle}) \right]$$

Vectorization.

$$\frac{\partial E}{\partial x^{\langle t \rangle}} = \frac{\partial E}{\partial p^{\langle t \rangle}}[n_c:, :] + \frac{\partial E}{\partial z^{\langle t \rangle}}[n_c:, :] + \frac{\partial E}{\partial q^{\langle t \rangle}}[n_c:, :]$$

$$\frac{\partial E}{\partial c^{\langle t-1 \rangle}} = \frac{\partial E}{\partial c^{\langle t \rangle}} * (1 - \Gamma_u^{\langle t \rangle}) + \frac{\partial E}{\partial J^{\langle t \rangle}} * \Gamma_r^{\langle t \rangle} + \frac{\partial E}{\partial z^{\langle t \rangle}}[:n_c, :]$$