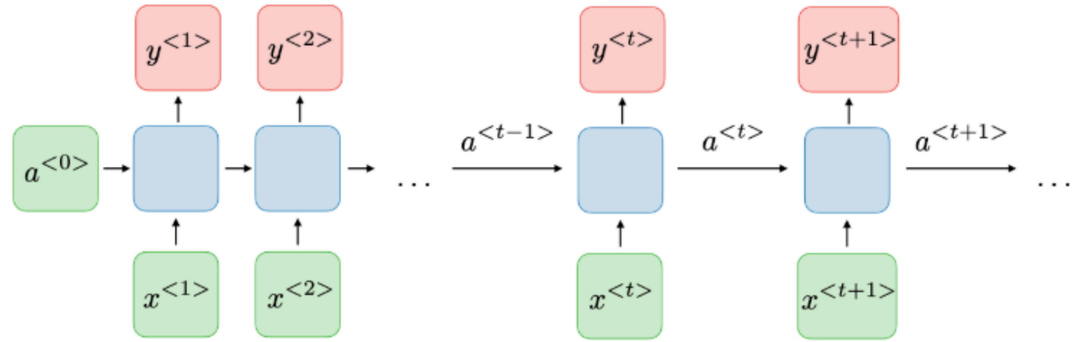
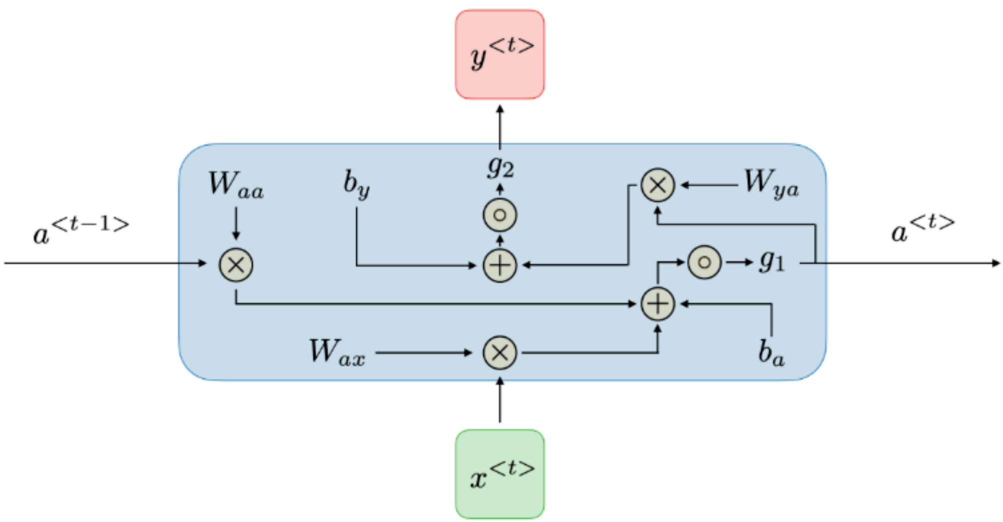




$$\boxed{a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)}$$ and $$\boxed{y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)}$$

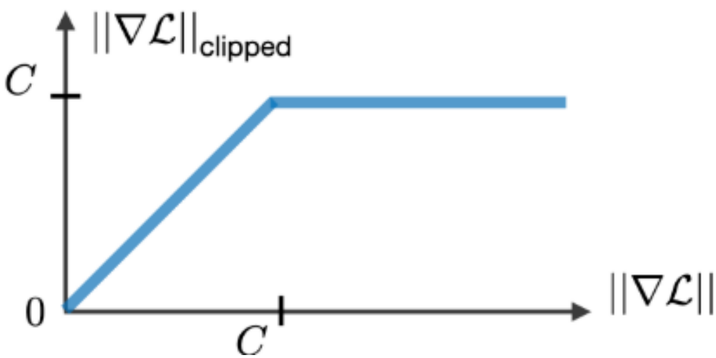
$$u_t = W_{aa}h_{t-1}+W_{ax}x_t+b_a,\, h_t = \tanh(u_t),\, o_t = W_{ya}h_t+b_y,\, \hat{y}_t = \mathrm{softmax}(o_t),\, L_t = -\sum_{i=1}^{K} y_{t,i}\log(\hat{y}_{t,i}),\, \mathcal{L} = \sum_{t=1}^{T} L_t$$

$$\delta_t^y \equiv \frac{\partial L_t}{\partial o_t} = \hat{y}_t - y_t,\, \delta_{T+1}^u \equiv 0,\, \delta_t^h \equiv \frac{\partial \mathcal{L}}{\partial h_t} = W_{ya}^\top \delta_t^y + W_{aa}^\top \delta_{t+1}^u,\, \delta_t^u \equiv \frac{\partial \mathcal{L}}{\partial u_t} = (1-h_t^2)\odot\delta_t^h.$$

$$\frac{\partial \mathcal{L}}{\partial W_{ya}} = \sum_{t=1}^{T}\delta_t^y h_t^\top,\, \frac{\partial \mathcal{L}}{\partial b_y} = \sum_{t=1}^{T}\delta_t^y,\, \frac{\partial \mathcal{L}}{\partial W_{aa}} = \sum_{t=1}^{T}\delta_t^u h_{t-1}^\top,\, \frac{\partial \mathcal{L}}{\partial W_{ax}} = \sum_{t=1}^{T}\delta_t^u x_t^\top,\, \frac{\partial \mathcal{L}}{\partial b_a} = \sum_{t=1}^{T}\delta_t^u.$$

$$W_{ya} \leftarrow W_{ya} - \eta\frac{\partial \mathcal{L}}{\partial W_{ya}},\, b_y \leftarrow b_y - \eta\frac{\partial \mathcal{L}}{\partial b_y},\, W_{aa} \leftarrow W_{aa} - \eta\frac{\partial \mathcal{L}}{\partial W_{aa}},\, W_{ax} \leftarrow W_{ax} - \eta\frac{\partial \mathcal{L}}{\partial W_{ax}},\, b_a \leftarrow b_a - \eta\frac{\partial \mathcal{L}}{\partial b_a}.$$

- **Gradient Clipping**



$$g_{clipped} = g\,\min\left(1, \frac{C}{\|g\|}\right)$$