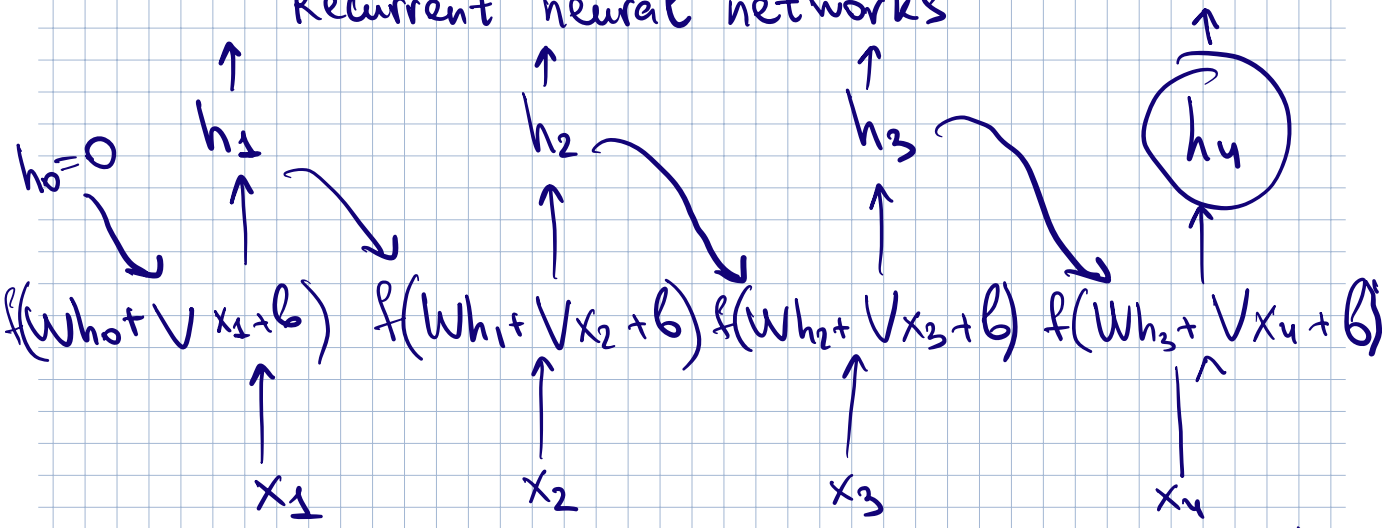
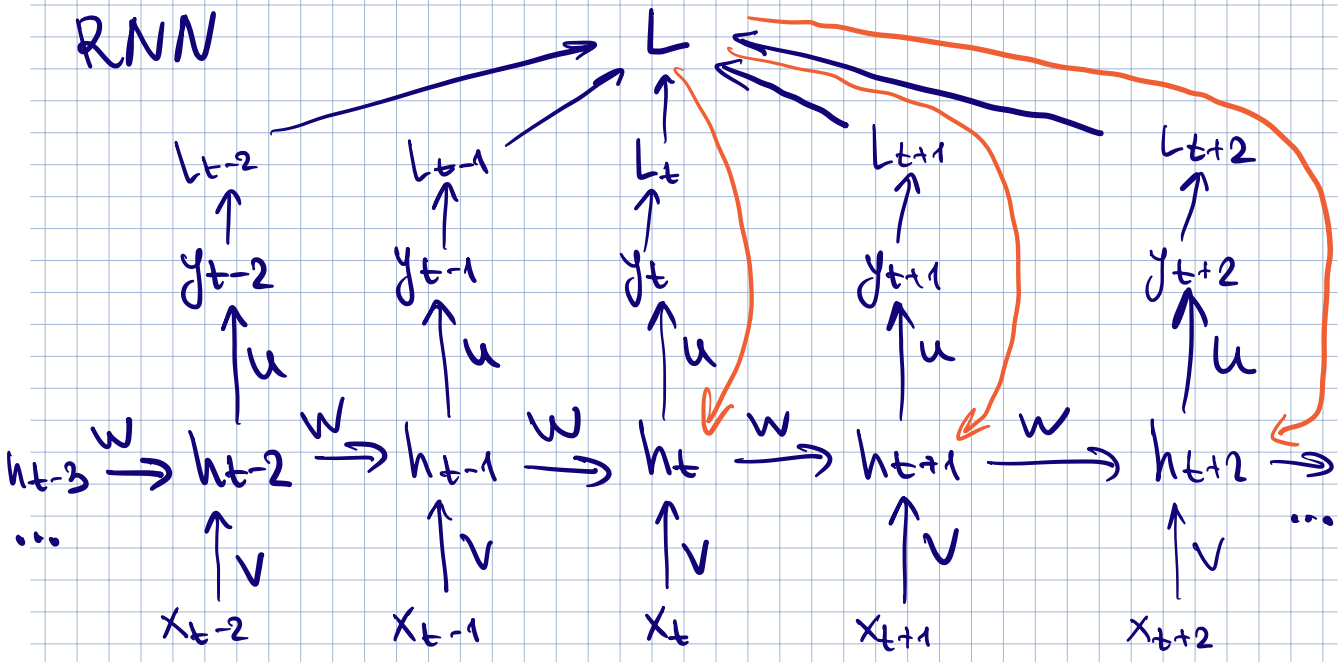
Recurrent neural networks



$x_i \in \mathbb{R}^d$ $\quad (V\ W)\begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$ $\quad h_t = f(Wh_{t-1} + Vx_t + b)$

RNN



- $L = \frac{1}{T}\sum_{t=1}^{T} L_t$
- $\frac{dL}{dU} = \frac{1}{T}\sum_{t=1}^{T} \frac{dL_t}{dU} = \frac{1}{T}\sum_{t=1}^{T} \frac{dL_t}{dy_t} \cdot \frac{dy_t}{dU}$

- $\frac{dL}{dW} = \frac{1}{T}\sum_{t=1}^{T}\frac{dL_t}{dW} = \frac{1}{T}\sum_{t=1}^{T}\frac{dL_t}{dy_t}\cdot\frac{dy_t}{dh_t}\cdot\frac{dh_t}{dW}$

$h_t = f(Vx_t + Wh_{t-1} + b)$

$\frac{dh_t}{dW} = \frac{\partial h_t}{\partial W} + \frac{\partial h_t}{\partial h_{t-1}}\cdot\frac{dh_{t-1}}{dW} = \frac{\partial h_t}{\partial W} + \frac{\partial h_t}{\partial h_{t-1}}\cdot$

$\cdot\left(\frac{\partial h_{t-1}}{\partial W} + \frac{\partial h_{t-1}}{\partial h_{t-2}}\cdot\frac{dh_{t-2}}{dW}\right) =$

$= \frac{\partial h_t}{W} + \frac{\partial h_t}{\partial h_{t-1}}\cdot\frac{\partial h_{t-1}}{\partial W} + \frac{\partial h_t}{\partial h_{t-1}}\cdot\frac{\partial h_{t-1}}{\partial h_{t-2}}\cdot\frac{dh_{t-2}}{dW}$

$= \frac{\partial h_t}{W} + \sum_{j=1}^{t-1}\cdot\left(\prod_{k=t-j}^{t-1}\frac{\partial h_{k+1}}{\partial h_k}\right)\cdot\frac{\partial h_{t-j}}{\partial W} =$

$= \sum_{j=0}^{t-1}\left(\prod_{k=t-j}^{t-1}\frac{\partial h_{k+1}}{\partial h_k}\right)\cdot\frac{\partial h_{t-j}}{\partial W}$

$\frac{dL}{dW} = \frac{1}{T}\sum_{t=1}^{T}\frac{dL_t}{dy_t}\cdot\frac{dy_t}{dh_t}\left[\sum_{j=0}^{t-1}\left(\prod_{k=t-j}^{t-1}\frac{\partial h_{k+1}}{\partial h_k}\right)\frac{\partial h_{t-j}}{\partial W}\right]$

BPTT (Backpropagation Through Time)

- $\left\|\frac{\partial h_{k+1}}{\partial h_k}\right\| > 1$ - exploding gradients
divergence, NaNs

Gradient clipping

$g \leftarrow \min\left(1, \frac{thr}{\|grad\|}\right)\cdot grad$

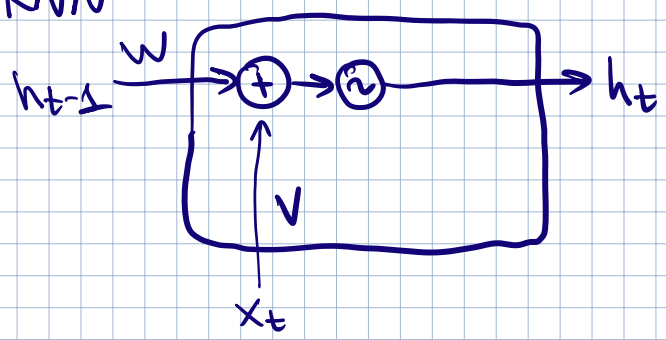
$$g \Leftarrow clip(grad, -c, c)$$

- $\left\| \frac{\partial h_{k+1}}{\partial h_k} \right\| < 1$ - vanishing gradients

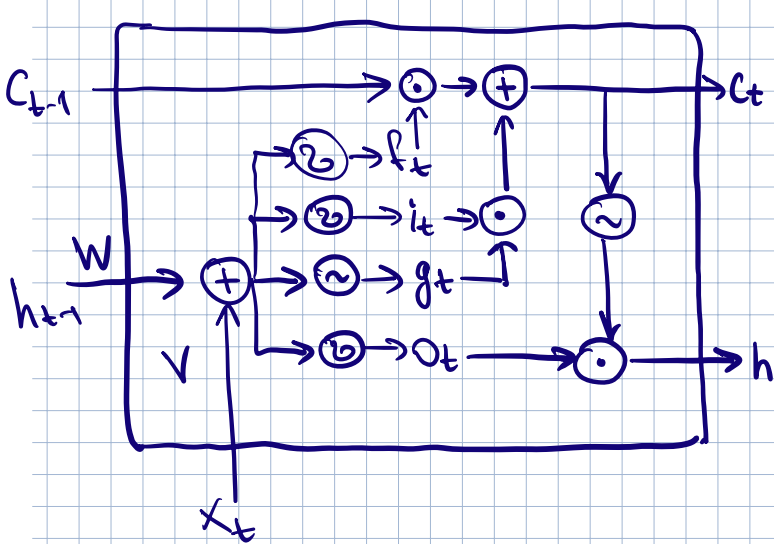
$$h_{k+1} = f(\overbrace{W h_k + V x_{k+1} + b}^{z_k})$$

$$\left\| \frac{\partial h_{k+1}}{\partial h_k} \right\| = \left\| diag(f'(z_k)) \right\| \cdot \|W\|$$ → init with ortogonal matrix

RNN

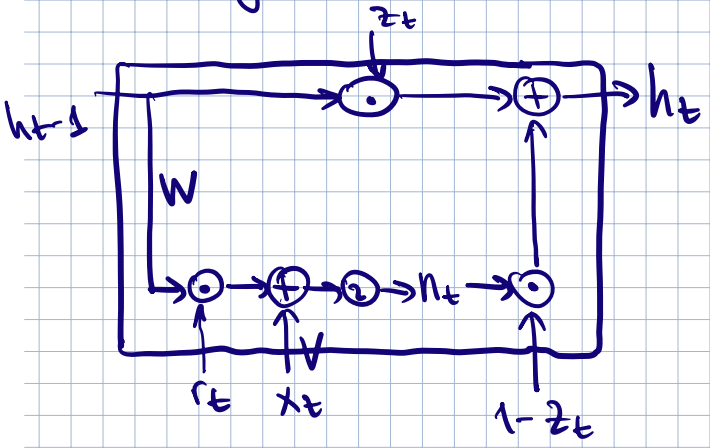


LSTM (long short-term memory)



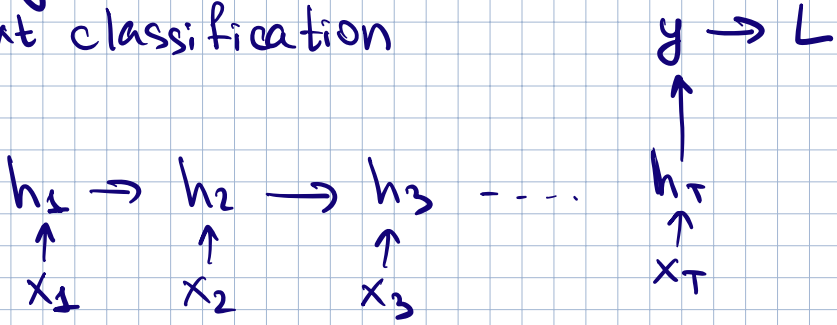
$$g_t = f(W_g h_{t-1} + V_g x_t + b_g)$$
$$i_t = \sigma(W_i h_{t-1} + V_i x_t + b_i)$$
$$o_t = \sigma(W_o h_{t-1} + V_o x_t + b_o)$$
$$f_t = \sigma(W_f h_{t-1} + V_f x_t + b_f)$$
$$C_t = C_{t-1} \odot f_t + i_t \odot g_t$$
$$h_t = f(C_t) \odot o_t$$

GRU (gated recurrent unit)



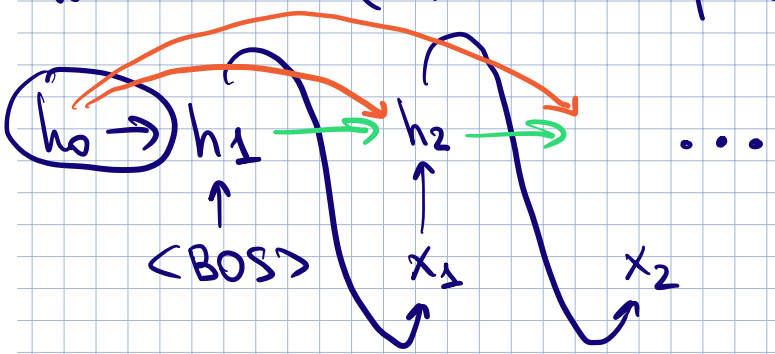
$$r_t = \sigma(V_r x_t + W_r h_{t-1} + b_r)$$
$$z_t = \sigma(V_z x_t + W_z h_{t-1} + b_z)$$
$$n_t = f(V_n x_t + r_t \odot (W_n h_{t-1} + b_{nk}) + b_n)$$
$$h_t = (1-z_t) \odot n_t + z_t \odot h_{t-1}$$

- Many-to-one
  - Text classification



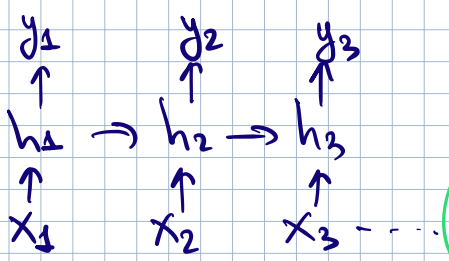
- One-to-many
  - Conditional generation (image captioning, genre music generation)

h_0 = Linear (Encoder (input))

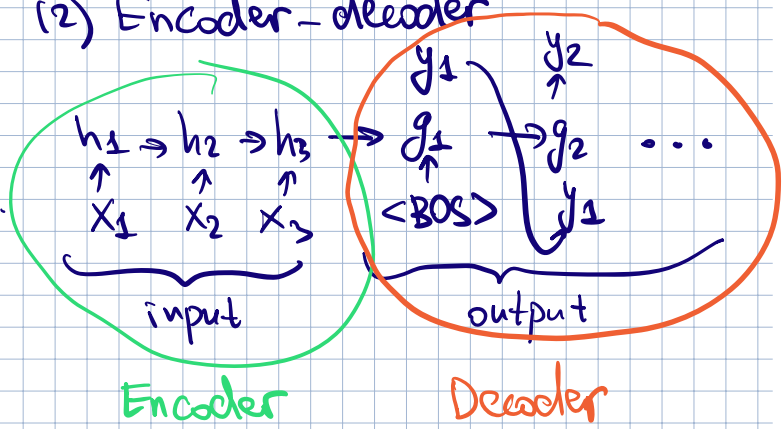


- Many-to-many (seq-to-seq)
  - Part of speech tagging (1)
  - Machine transalion (2)

(1)



(2) Encoder-decoder

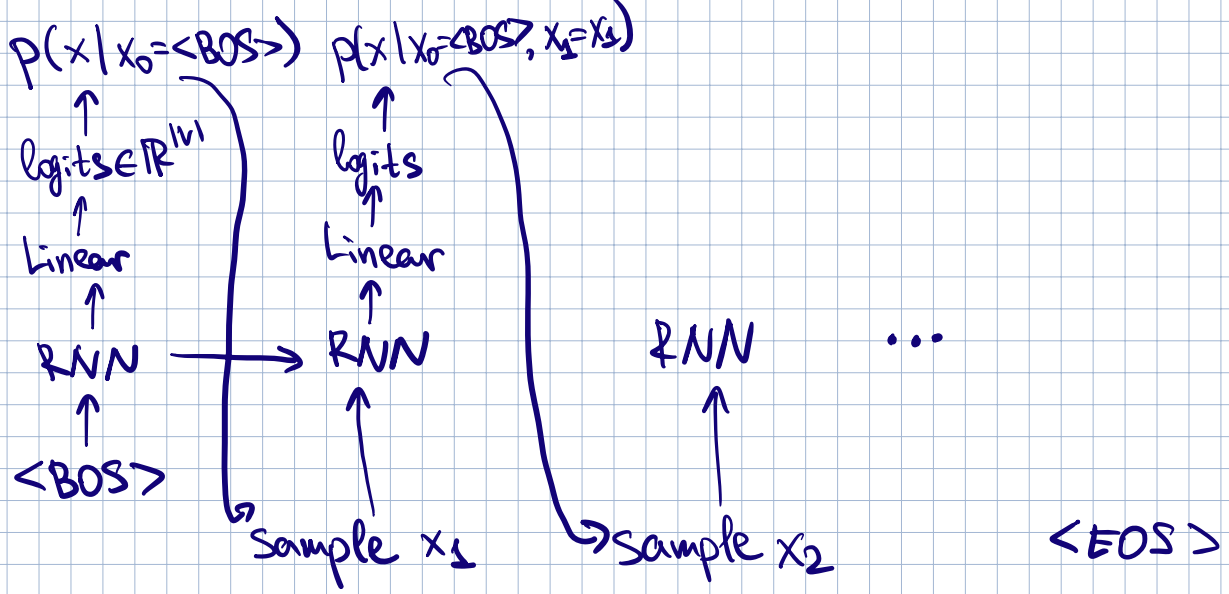


Language model

Autoregressive generation

$x = (x_0, x_1, x_2, \ldots, x_T)$

$p(x) = p(x_0)\, p(x_1|x_0) \cdot p(x_2|x_0, x_1) \cdot \ldots \cdot p(x_T|x_0, \ldots, x_{T-1})$



Θ - model parameters

$p_\Theta(x_t \mid x_0, x_1, \ldots, x_{t-1})$

$X = (x_0, \ldots, x_T)$ - real sequence from dataset

$$p_\Theta(x) = \underbrace{p(x_0)}_{=1} \cdot p_\Theta(x_1 \mid x_0) \cdot \ldots \cdot p(x_T \mid x_0, \ldots, x_{T-1})$$

$$-\sum_{t=1}^{T} \log p_\Theta(x_t \mid x_0, \ldots, x_{t-1}) \to \min_\Theta$$