a language model that can handle unbounded context (prefix).

$$p(x_1, x_2, \cdots, x_T) = \prod_{t=1}^{T} \underbrace{p(x_t | x_1, \cdots, x_{t-1})}_{\approx\, p(x_t | x_{t-n}, \cdots, x_{t-1}) \text{ : n-gram LM}}$$

$x_t \in V$

Neural probabilistic language model.

① $\underbrace{e(x_{t-n}, \cdots, x_{t-1})} = [e(x_{t-n})^T ; \cdots , e(x_{t-1})^T]^T \in \mathbb{R}^{nd}$

② $h = \tanh(W e(x_{t-n}, \cdots, x_{t-1}) + b) \in \mathbb{R}^{d'}$ $W \in \mathbb{R}^{d' \times nd}$

③ $F(x_t, x_{t-n,t-1}) = \underset{\text{class embedding}}{e'(x_t)}^T h + b'(x_t)$

$W \in \mathbb{R}^{d' \times nd}$

$$d' \left\{ \begin{bmatrix} \underbrace{W_{t-n}}_{d} & \underbrace{W_{t-n+1}}_{d} & \cdots & \underbrace{W_{t-1}}_{d} \end{bmatrix} \right. \begin{bmatrix} e(x_{t-n}) \\ e(x_{t-n+1}) \\ \vdots \\ e(x_{t-1}) \end{bmatrix} = \underline{\sum_{n'=1}^{n} W_{t-n'} e(x_{t-n'})}$$

A text classifier summarizes the input text

$X = (x_1, \cdots, x_T)$

$\phi(X) = \sigma(W \sum_{t=1}^{T} e(x_t) + b)$, $F(X, y) = u_y^T \phi(X) + c_y$

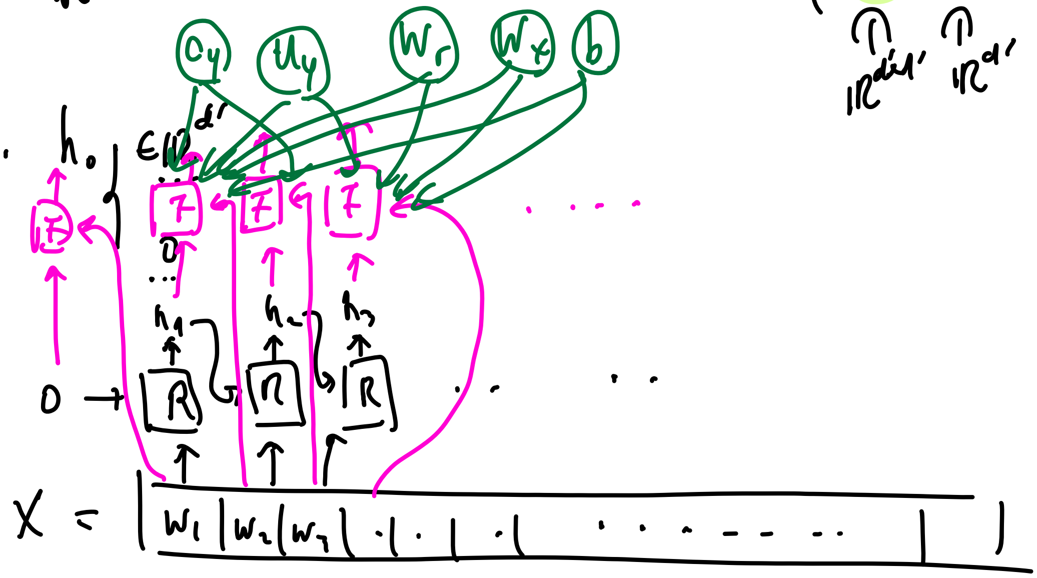
A recurrent neural network.

: recursively compress the input sequence.

If $h_{t-1} = \phi(X=\{w_1,\cdots,w_{t-1}\})$, can we add w_t?

. $\underset{\mathbb{R}^{d'}}{\underset{\cap}{h_t}} = R(\underset{\mathbb{R}^{d'}}{\underset{\cap}{h_{t-1}}}, w_t)$, where e.g. $R(h_{t-1}, w_t)$

$$= \sigma(\underset{\mathbb{R}^{d'\times d'}}{\underset{\cap}{W_r}}\, \underset{\mathbb{R}^{d'}}{\underset{\cap}{h_{t-1}}} + \underset{\mathbb{R}^{d'\times d}}{\underset{\cap}{W_x}}\, \underset{\mathbb{R}^{d}}{\underset{\cap}{e(w_t)}} + \underset{\mathbb{R}^{d'}}{\underset{\cap}{b}})$$

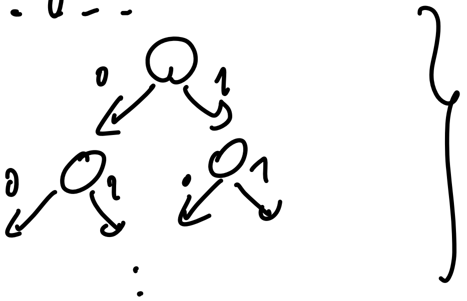


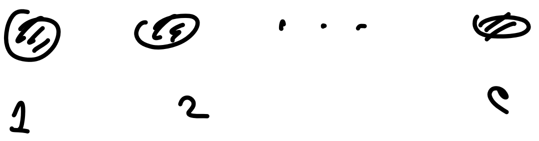
A recurrent neural net language model [Mikolov et al. 2010]

Alice ⟹ Bob

$m \in \{1,\cdots,c\}$

$\log_2 c$ bits





$$-\sum p(m) \log_2 p(m)$$

$$2^{-\log_2 p(m)} \Rightarrow \boxed{\underline{2^{-\log_2 p(w_t | w_{<t})}}}$$

$$B = -\frac{1}{\sum_{n=1}^{N} T^n} \sum_{n=1}^{N} \sum_{t=1}^{T^n} \log_2 p(w_t^n | w_{<t}^n)$$

...

, perplexity $= 2^{\beta}$