LSTM (h_t, C_t)

short $\quad$ long.

① Outcome depends on short-term

$$o_t = \sigma\left(W_o \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} + b_o\right)$$ output gate

$W_o \in \mathbb{R}^{\underline{d_c} \times (\underline{d_h} + d_x)}$

the same,

② Short-term state Depends on long term state

$$h_t = o_t * \tanh(C_t)$$

$o_t \in \mathbb{R}^{d_c}$, $\quad C_t \in \mathbb{R}^{d_c}$

$h_t \rightarrow$ produce a distribution on V.

③ long-term states depend on short-term state & previous periods

$$C_t = \underbrace{f_t}_{\text{forget gate}} * C_{t-1} + \underbrace{i_t}_{\text{input gate}} * \tilde{C}_t$$

④ $$f_t = \sigma\left(W_f \cdot \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} + b_f\right)$$

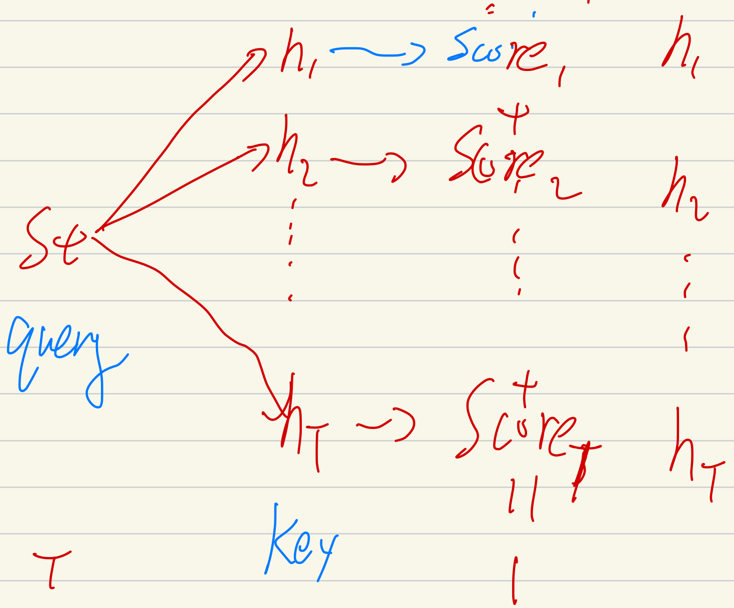
⑤ $$i_t = \sigma\left(W_i \cdot \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} + b_i\right)$$

$$\tilde{C}_t = \tanh\left(W_c \cdot \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} + b_c\right)$$

Attention Mechanism

Encoder $\qquad\qquad\qquad\qquad$ Decoder

$$h_1 \to h_2 \to h_3 \to \cdots \to h_T \qquad\qquad s_t$$



$$a_t = \sum_{j=1}^{T} \text{score}_j \cdot h_j$$

$a_t \in \mathbb{R}^d$

why is Attention great?

$$\text{score} = \text{softmax}(e)$$

$\text{score} \in \mathbb{R}^T$

$$e_i = h_i^T \cdot s_t \in \mathbb{R}$$

$$\begin{pmatrix} a_t \\ s_t \end{pmatrix}$$

$x_1, x_2, \cdots, x_n \in \mathbb{R}^d$

$W_q, W_k, W_v \in \mathbb{R}^{d\times d}$

$\in \mathbb{R}^d$

$q_i = W_q \cdot x_i, \quad k_i = W_k \cdot x_i, \quad v_i = W_v \cdot x_i$

query key Value

w'

$$w'_{ij} = \frac{q_i \cdot k_j}{\sqrt{d}} \qquad i, j \in [n]$$

$\downarrow$ Softmax

$$w_{ij} = \frac{\exp(w'_{ij})}{\sum_{j'=1}^{n} \exp(w'_{i,j'})}$$

$$y_i = \sum_{j=1}^{n} w_{ij} v_j \in \mathbb{R}^d \qquad \text{linear in } v.$$

How can we do better?

(1) Better parallelization?

Multi-head attention.

(2) non-linear v?

Add an MLP layer.

(3) How to keep the sequence information?

Position encoding/embedding.

$$\text{input} = \vec{X} + \text{Position Encoding}$$

(4) No future information? / Auto-regressive

masked attention

(5) Can we pray to Optimization God better?

(i) Skip-Connection

$f(x) + x$

(2) Layer Normalization