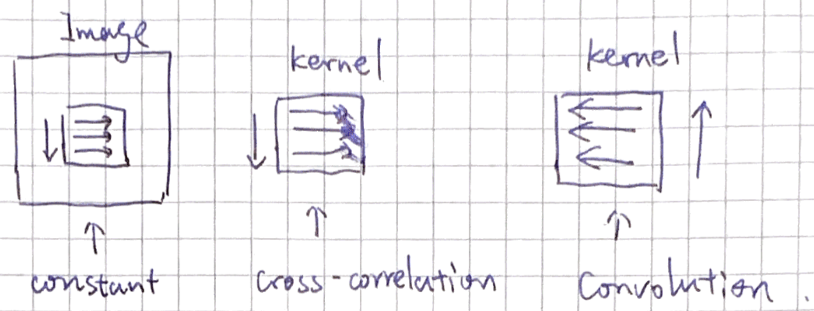
Updates:

(1) $z = x^T W^T \Rightarrow J_z = \frac{\partial z}{\partial x}$, Matrix differentiation, P_{13}

(2) Convolutions v.s. Corss-correlation



(3) sonlution will be uploaded after tatorial session.

(4) CIFAR10. Competetion. 95% ~ 96% word record.

(1) $$z = (W_x)^T = \left[\begin{pmatrix} W_{11} & \cdots & W_{1d} \\ \vdots & & \\ W_{n1} & \cdots & W_{nd} \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}\right]^T = \left(\sum_{i=1}^{d} W_{1i} x_i, \cdots \sum_{i=1}^{d} W_{ni} x_i\right).$$

$\frac{\partial z}{\partial x}$ ← row vector, ↖ colum vector

$$\frac{\partial z}{\partial x} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_n}{\partial x_1} \\ \vdots & & \\ \frac{\partial z_1}{\partial x_d} & \cdots & \frac{\partial z_n}{\partial x_d} \end{bmatrix} \Big\} d = W^T.$$

$\underbrace{\qquad}_{n}$ dxn matrix

Ex 5-1 (a) $y \in \mathbb{R}^n$, $x \in \mathbb{R}^d$ $W = \mathbb{R}^{n \times d}$.

$y = \sigma(W_x) \Rightarrow \frac{\partial y}{\partial x}$? $\quad \frac{\partial y_i}{\partial x_j} = \frac{\partial \sigma(W_i^T x)}{\partial x_j} = \sigma'(W_i^T x) \cdot \frac{\partial W_i^T x}{\partial x_j}$

$y_i = \sigma(W_i^T x)$

$$= \sigma'(\dots) \frac{\partial \sum_{k=1}^{d} W_{ik} x_k}{\partial x_j} = \sigma'(\dots) W_{ij}.$$

$$= \sigma'(W_i^T x) W_{ij}.$$

$$\therefore \frac{\partial y}{\partial x} = \begin{pmatrix} \sigma'(W_1^T x) W_{11} & \cdots & \sigma'(W_1^T x) \cdot W_{1d} \\ \vdots & & \\ \sigma'(W_n^T x) W_{n1} & \cdots & \sigma'(W_n^T x \cdot W_{nd} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma'(W_1^T x) & & 0 \\ & \ddots & \\ 0 & & \sigma'(W_n^T x) \end{pmatrix}\begin{pmatrix} W_{11} & \cdots & W_{1d} \\ \vdots & & \vdots \\ W_{n1} & \cdots & W_{nd} \end{pmatrix} = \mathrm{diag}(\sigma') \cdot W.$$

(b) $\frac{\partial L}{\partial W} = \frac{\partial}{\partial W}\sum_{t=0}^{T} L_t = \sum_{t=0}^{T}\frac{\partial L_t}{\partial W} = \sum_{t=0}^{T}\frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial W}$.

h_t depends on h_k $(k \le t)$. h_k depends on W.

$$\sum_{t=0}^{T}\frac{\partial L_t}{\partial h_t}\cdot\frac{\partial h_t}{\partial W} = \sum_{t=0}^{T}\sum_{k=0}^{t}\underbrace{\frac{\partial L_t}{\partial h_t}}_{\text{unknow}}\cdot\frac{\partial h_t}{\partial h_k}\cdot\underbrace{\frac{\partial h_k}{\partial W}}_{\text{trival}}$$

$$\frac{\partial h_t}{\partial h_k} = \prod_{i=k+1}^{t}\frac{\partial h_i}{\partial h_{i-1}} = \frac{\partial h_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial h_{t-2}}\cdots\frac{\partial h_{k+1}}{\partial h_k}.$$

$h_i = \sigma(W h_{i-1} + W_i x)$

$$\Rightarrow \frac{\partial L}{\partial W} = \sum_{t=0}^{T}\sum_{k=0}^{t}\frac{\partial L_t}{\partial h_t}\cdot\prod_{i=k+1}^{t}\frac{\partial h_i}{\partial h_{i-1}}\cdot\frac{\partial h_k}{\partial W}.$$

$\hookrightarrow \prod_{i=1}^{t}[\mathrm{diag}(\sigma'(W_i^T x))W]$

Ex 5-2.

$T=3$ $\quad t = 0, 1, 2, 3$.

(a) $\frac{\partial L}{\partial W} = \frac{\partial L_0}{\partial h_0}\cdot\frac{\partial h_0}{\partial h_0}\frac{\partial h_0}{\partial W}$ $\quad(t=0)$

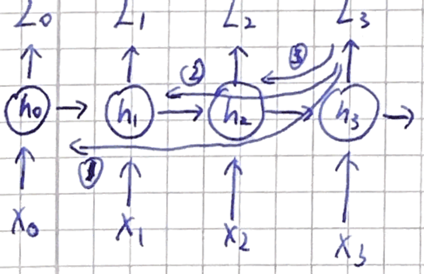
$+ \frac{\partial L_1}{\partial h_1}\frac{\partial h_1}{\partial h_0}\cdot\frac{\partial h_0}{\partial W} + \frac{\partial L_1}{\partial h_1}\cdot\frac{\partial h_1}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}$ $\quad(t=1)$

$(t=2)$

$+ \frac{\partial L_2}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_0}\cdot\frac{\partial h_0}{W} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial h_2}\cdot\frac{\partial h_2}{\partial h_0}\cdot\frac{\partial h_0}{\partial W}$.

$+ \underbrace{\frac{\partial L_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial h_0}\cdot\frac{\partial h_0}{\partial W}}_{①} + \underbrace{\frac{\partial L_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}}_{②} + \underbrace{\frac{\partial h_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}}_{③} + \frac{\partial L_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial W}$

$(t=3)$



$= \frac{\partial L_0}{\partial h_0}\cdot\frac{\partial h_0}{\partial W}$.

$+ \frac{\partial L_1}{\partial h_1}\cdot \mathrm{diag}(\sigma')W\cdot\frac{\partial h_0}{\partial W} + \frac{\partial L_1}{\partial h_1}\cdot\frac{\partial h_1}{\partial W}$

matrix product $+ \frac{\partial L_2}{\partial h_2}\cdot[\mathrm{diag}(\sigma')W]^2\frac{\partial h_0}{\partial W} + \frac{\partial L_2}{\partial h_2}\mathrm{diag}(\sigma')W\cdot\frac{\partial h_0}{\partial W}$.

$+ \frac{\partial L_3}{\partial h_3}[\mathrm{diag}(\sigma')W]^3\frac{\partial h_0}{\partial W} + \frac{\partial L_3}{\partial h_3}[\mathrm{diag}(\sigma')W]^2\frac{\partial h_0}{\partial W} + \dots$ $\quad + \frac{\partial L_2}{\partial h_2}\frac{\partial h_1}{\partial W}$

Ex 5-2 . (b)

$$M = Q\Lambda Q^{-1} \qquad M^n = Q\Lambda^n Q^{-1}.$$

Let $Q = (Q_1 \cdots, Q_n)$

$$MQ = M(Q_1, \cdots, Q_n) = (MQ_1, \cdots, MQ_n)$$

$$\underline{\underline{\text{eigen equation}}} \quad (\lambda_1 Q_1, \cdots, \lambda_n Q_n)$$

$$= (Q_1, \cdots, Q_n) \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} = Q\Lambda.$$

Thus: $M = Q\Lambda Q^{-1}$

Hypothesis: $M^k = Q\Lambda^k Q^{-1}$.

$$M^{k+1} = MM^k = Q\Lambda Q^{-1}(Q\Lambda^k Q^{-1})$$

$$= Q\Lambda Q^{-1} Q\Lambda^k Q^{-1} = Q\Lambda \cdot \Lambda^k Q^{-1} = Q\Lambda^{k+1} Q^{-1}.$$

#

(c) $\prod_{i=1}^{t} diag(\sigma')W = diag(\sigma')^n \cdot W^n$.

$$W^n = Q\Lambda^n Q^{-1} \Rightarrow \Lambda^{\circ n} = \begin{pmatrix} \lambda_1^n & & 0 \\ & \ddots & \\ 0 & & \lambda_n^n \end{pmatrix}$$

$$\lim_{n\to\infty} \lambda_i^n = \begin{cases} \infty & \text{if } \lambda_i > 1 \\ 1 & \text{if } \lambda_i = 1 \\ 0 & \text{else } \lambda_i < 1 \end{cases}.$$

Ex 5-3

(a) Input gate: Selectively updates the cell state based on new inputs.

Output gate: output is the filtered version of the cell state.

Forget gate: - decides what information to throw away or remember from previous cell state.

- decision maker: sigmoid layer (forget gate layer.

(b) $f_t, i_t, o_t = \sigma(\cdots)$ positive.

(c) $\frac{\partial C_t}{\partial C_{t-1}} \approx 1$

$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

$\tilde{C}_t = \tanh(W_c h_{t-1} + U_c x_t)$

$$\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial}{\partial C_{t-1}}\left(f_t \odot C_{t-1} + i_t \cdot \tanh(W_c h_{t-1} + U_c x_t)\right)$$

$h_{t-1} \to C_{t-1}$

$$= \frac{\partial f_t}{\partial C_{t-1}} \odot C_{t-1} + f_t \cdot \frac{\partial C_{t-1}}{\partial C_{t-1}} + \frac{\partial i_t}{\partial C_{t-1}} \odot \tilde{C}_t + i_t \odot \frac{\partial \tilde{C}_t}{\partial C_{t-1}}$$

$$= \frac{\partial f}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial C_{t-1}} \odot C_{t-1} + f_t + \frac{\partial i_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}} \odot \tilde{C}_t + i_t \odot \frac{\partial \tilde{C}_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}}$$

$$= \sigma' W_f \cdot \left[\frac{\partial O_{t-1}}{\partial C_{t-1}} \odot C_{t-1} + \tanh'()\right] \odot C_{t-1}$$

$$+ \sigma' W_i \left[\frac{\partial O_{t-1}}{\partial C_{t-1}} \odot C_{t-1} + \tanh'()\right] \odot \tilde{C}_t$$

$$+ i_t \odot \tanh'() \left[\frac{\partial O_{t-1}}{\partial C_{t-1}} \odot C_{t-1} + \tanh'()\right]$$

$$= \sigma' W_f \delta \odot C_{t-1} + f_t + \sigma' W_i \delta \odot \hat{C}_t + i_t \odot \tanh'() \delta$$

$\delta = \frac{\partial O_{t-1}}{\partial C_{t-1}} \odot C_{t-1} + \tanh$

$f_t = \sigma(W_f h_{t-1} + U_f x_t)$

$i_t = \sigma(W_i h_{t-1} + U_i x_t)$

$O_t = \sigma(W_o h_{t-1} + U_o x_t$

$\tilde{C}_t = \tanh(W_c h_{t-1} + U_c x_t$

$h_t = O_t \odot \tanh(C_t)$