1. $w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \in \mathbb{R}^{2\times3\times3}$, $w_1, w_2 \in \mathbb{R}^{3\times3}$

$$w_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad w_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

□

2. AvgPool2d can be represented as a convolution

with a filter $w = \begin{bmatrix} w_1 \\ \vdots \\ w_C \end{bmatrix} \in \mathbb{R}^{C\times C\times K\times K}$

$$w_i = \left[\underline{0}, \cdots, \underline{0}, \begin{bmatrix} 1/K^2 & \cdots & 1/K^2 \\ \vdots & \ddots & \vdots \\ 1/K^2 & \cdots & 1/K^2 \end{bmatrix}, \underline{0}, \cdots, \underline{0} \right]^T \in \mathbb{R}^{C\times K\times K}, \quad i = 1, \cdots, C$$

→ i-th element of w_i

$\underline{0}$: $K\times K$ zero matrix

with stride K without zero paddings.

□

3. $w = \begin{bmatrix} [[0.299]], \\ [[0.587]], \\ [[0.114]] \end{bmatrix} \in \mathbb{R}^{3\times1\times1}$ be the filter.

4. Since $\sigma : \mathbb{R} \to \mathbb{R}$ nondecreasing,

$x_1 \geq x_2 \Rightarrow \sigma(x_1) \geq \sigma(x_2)$, $\forall x_1, x_2 \in \mathbb{R}$

$\therefore \rho(x) = x_{ij} \Rightarrow \rho(\sigma(x)) = \sigma(x_{ij})$, $x \in \mathbb{R}^{p\times p}$, for some i, j

Where p is the kernel size of max pool ρ.

That is, $\sigma(\rho(x)) = \sigma(x_{ij}) = \rho(\sigma(x))$.

Since stride of max pooling ρ is also p,

the equation $\sigma(\rho(x)) = \rho(\sigma(x))$ holds for $\forall x \in \mathbb{R}^{m\times n}$

□

6.

(a) $\frac{\partial y_L}{\partial b_L} = \frac{\partial}{\partial b_L}\left(\cancel{A_L y_{L-1}} + b_L\right) = \frac{\partial}{\partial b_L}\begin{bmatrix}(b_L)_1 \\ \vdots \\ (b_L)_{n_L}\end{bmatrix} = I$

$\frac{\partial y_L}{\partial y_{L-1}} = \frac{\partial}{\partial y_{L-1}}\left(A_L y_{L-1} + \cancel{b_L}\right) = \frac{\partial}{\partial y_{L-1}}\left[(A_L)_{11}(y_{L-1})_1 + \cdots + (A_L)_{1,n_{L-1}}(y_{L-1})_{n_{L-1}}\right]$

$= \begin{bmatrix}(A_L)_{11} & \cdots & (A_L)_{1,n_{L-1}}\end{bmatrix} = A_L$

$\frac{\partial y_l}{\partial b_l} = \frac{\partial}{\partial b_l}\sigma(A_l y_{l-1} + b_l) = \frac{\partial}{\partial b_l}\begin{bmatrix}\sigma\left((A_l)_{11}(y_{l-1})_1 + \cdots + (A_l)_{1,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_1\right) \\ \vdots \\ \sigma\left((A_l)_{n_l 1}(y_{l-1})_1 + \cdots + (A_l)_{n_l,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_{n_l}\right)\end{bmatrix}$

$= \begin{bmatrix}\sigma'\left((A_l)_{11}(y_{l-1})_1 + \cdots + (A_l)_{1,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_1\right) & & 0 \\ & \ddots & \\ 0 & & \sigma'\left((A_l)_{n_l,1}(y_{l-1})_1 + \cdots + (A_l)_{n_l,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_{n_l}\right)\end{bmatrix}$

$= \operatorname{diag}\left(\sigma'(A_l y_{l-1} + b_l)\right), \quad l = 1, \cdots, L-1$

$\frac{\partial y_l}{\partial y_{l-1}} = \frac{\partial}{\partial y_{l-1}}\sigma(A_l y_{l-1} + b_l) = \frac{\partial}{\partial y_{l-1}}\begin{bmatrix}\sigma\left((A_l)_{11}(y_{l-1})_1 + \cdots + (A_l)_{1,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_1\right) \\ \vdots \\ \sigma\left((A_l)_{n_l 1}(y_{l-1})_1 + \cdots + (A_l)_{n_l,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_{n_l}\right)\end{bmatrix}$

$= \begin{bmatrix}\sigma'\left((A_l)_{11}(y_{l-1})_1 + \cdots + (A_l)_{1,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_1\right) & & 0 \\ & \ddots & \\ 0 & & \sigma'\left((A_l)_{n_l,1}(y_{l-1})_1 + \cdots + (A_l)_{n_l,n_{l-1}}(y_{l-1})_{n_{l-1}} + (b_l)_{n_l}\right)\end{bmatrix}\begin{bmatrix}(A_l)_{11} & \cdots & (A_l)_{1,n_{l-1}} \\ & \ddots & \\ (A_l)_{n_l,1} & \cdots & (A_l)_{n_l,n_{l-1}}\end{bmatrix}$

$= \operatorname{diag}\left(\sigma'(A_l y_{l-1} + b_l)\right) A_l, \quad l = 2, \cdots, L-1$

(b) $$\left(\frac{\partial y_L}{\partial A_L}\right)_{ij} = \frac{\partial y_L}{\partial (A_L)_{ij}} = \frac{\partial}{\partial (A_L)_{ij}}\left(A_L y_{L-1} + \cancel{b_L}\right)$$

$$= \frac{\partial}{\partial (A_L)_{ij}}\left((A_L)_{i1}(y_{L-1})_1 + \cdots + (A_L)_{in_{L-1}}(y_{L-1})_{n_{L-1}}\right)$$

$$= (y_{L-1})_j$$

$$\therefore \frac{\partial y_L}{\partial A_L} = y_{L-1}^T$$

$$\left(\frac{\partial y_L}{\partial A_\ell}\right)_{ij} = \frac{\partial y_L}{\partial (A_\ell)_{ij}} = \frac{\partial y_L}{\partial y_\ell}\frac{\partial y_\ell}{\partial (A_\ell)_{ij}}$$

$$= \frac{\partial y_L}{\partial y_\ell}\frac{\partial}{\partial (A_\ell)_{ij}}\sigma(A_\ell y_{\ell-1} + b_\ell)$$

$$= \frac{\partial y_L}{\partial y_\ell}\frac{\partial}{\partial (A_\ell)_{ij}}\begin{bmatrix} \sigma\left((A_\ell)_{1,1}(y_{\ell-1})_1 + \cdots + (A_\ell)_{1,n_{\ell-1}}(y_{\ell-1})_{n_{\ell-1}} + (b_\ell)_1\right) \\ \vdots \\ \sigma\left((A_\ell)_{n_\ell,1}(y_{\ell-1})_1 + \cdots + (A_\ell)_{n_\ell,n_{\ell-1}}(y_{\ell-1})_{n_{\ell-1}} + (b_\ell)_{n_\ell}\right) \end{bmatrix}$$

$$= \frac{\partial y_L}{\partial y_\ell}\begin{bmatrix} 0 \\ \sigma'\left((A_\ell)_{i1}(y_{\ell-1})_1 + \cdots + (A_\ell)_{in_{\ell-1}}(y_{\ell-1})_{n_{\ell-1}} + (b_\ell)_i\right)\cdot (y_{\ell-1})_j \\ 0 \end{bmatrix}$$

$$= \sigma'\left((A_\ell)_{i1}(y_{\ell-1})_1 + \cdots + (A_\ell)_{in_{\ell-1}}(y_{\ell-1})_{n_{\ell-1}} + (b_\ell)_i\right)\left(\frac{\partial y_L}{\partial y_\ell}\right)_i (y_{\ell-1})_j$$

$$\therefore \frac{\partial y_L}{\partial A_\ell} = \operatorname{diag}\left(\sigma'(A_\ell y_{\ell-1} + b_\ell)\right)\left(\frac{\partial y_L}{\partial y_\ell}\right)^T y_{\ell-1}^T$$

□