#1

If we set $w = \left[ \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \right]$, $w_1$ will work as creating $Y_1$ and $w_2$ will work as creating $Y_2$.

Since we have zero-paddings augment the matrix $X$ such that the index starts at 0. $(X_{0,:} = X_{:,0} = 0)$

$$\begin{cases} Y_{1,i,j} = \sum_{a=1}^{3} \sum_{b=1}^{3} w_{1,(i-1)+a,(j-1)+b} X_{(i-2)+a,(j-2)+b} = -X_{ij} + X_{i+1,j} & \text{for } 1 \le i,j \le n \\ Y_{2,i,j} = \sum_{a=1}^{3} \sum_{b=1}^{3} w_{2,(i-1)+a,(j-1)+b} X_{(i-2)+a,(j-2)+b} = -X_{ij} + X_{i,(j+1)} \end{cases}$$

$\blacksquare$

#2.

If we consider the filter $w \in \mathbb{R}^{C \times C \times k \times k}$ s.t. $w_{xyzw} = \frac{1}{k^2} \delta_{xy}$ and no bias,

the output of the convolution operation of $X$ with $w$ with stride $k$ will be a tensor of size $C \times \frac{n}{k} \times \frac{n}{k}$.

Also $$Y_{cij} = \sum_{d=1}^{C} \sum_{a=1}^{k} \sum_{b=1}^{k} w_{c,d,k(i-1)+a,k(j-1)+b} \cdot X_{d,k(i-1)+a,k(j-1)+b}$$

$$= \sum_{d=1}^{C} \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{k^2} \delta_{cd} X_{d,k(i-1)+a,k(j-1)+b}$$

$$= \frac{1}{k^2} \sum_{a=1}^{k} \sum_{b=1}^{k} X_{c,k(i-1)+a,k(j-1)+b}$$

$\blacksquare$

#3

set the filter $w \in \mathbb{R}^{3 \times 1 \times 1}$ as $w_{1,1,1} = 0.299$, $w_{2,1,1} = 0.587$, $w_{3,1,1} = 0.114$, and set bias to False

Now operate the convolution operation of $X$ and $w$ with stride 1.

Then $$Y_{ij} = \sum_{c=1}^{3} \sum_{a=1}^{1} \sum_{b=1}^{1} w_{c,(i-1)+a,(j-1)+b} X_{c,(i-1)+a,(j-1)+b}$$

$$= w_{1,1,1} X_{1,i,j} + w_{2,1,1} X_{2,i,j} + w_{3,1,1} X_{3,i,j}$$

$$\therefore Y = 0.299 R + 0.587 G + 0.114 B.$$

$\blacksquare$

#4

If we perform $\rho(X)$, we take the maximum value inside the filter, move the filter, and repeat this process. Let's say one max-pooling step with a filter as a "step"

In each step, we chose the maximum value among the set of numbers determined in $X$ by the filter, denote it as $N$

For an arbitrary step, if $x = \max N$, the corresponding element for $\sigma(\rho(X))$ is $\sigma(x)$.

if we consider $\rho(\sigma(X))$, denote $\sigma(N) := \{\sigma(s) \mid s \in N\}$.

$\max(\sigma(N)) = \sigma(x)$ since $\sigma$ is a non-decreasing function.

(If $\max(\sigma(N)) = \sigma(y) > \sigma(x) \Rightarrow y > x$ which is a contradiction)

$\Rightarrow$ Since the element of each entry of $\sigma\rho(X)$ and $\rho(\sigma(X))$ are identical, they are the same.

$\blacksquare$

#6

(a) If $w = Av$, $\frac{\partial w}{\partial v} = A$

pf) $\left(\frac{\partial w}{\partial v}\right)_{ij} = \frac{\partial w_i}{\partial v_j} = A_{ij}$.

$\Rightarrow \frac{\partial y_L}{\partial b_L} = 1$, $\frac{\partial y_L}{\partial y_{L-1}} = A_L$. since $y_L \in \mathbb{R}$.

$y_\ell = \sigma(A_\ell y_{\ell-1} + b_\ell)$, $w_\ell := A_\ell y_{\ell-1} + b_\ell$

$$\Rightarrow \frac{\partial y_\ell}{\partial b_\ell} = \frac{\partial y_\ell}{\partial w_\ell}\cdot\frac{\partial w_\ell}{\partial b_\ell} = \frac{\partial y_\ell}{\partial w_\ell}\cdot I$$

$$\left(\frac{\partial y_\ell}{\partial w_\ell}\right)_{ij} = \frac{\partial (y_\ell)_i}{\partial (w_\ell)_j} = \frac{\partial}{\partial (w_\ell)_j}\left[\sigma((w_\ell)_i)\right] = \sigma'((w_\ell)_i)\cdot\delta_{ij}$$

$$\Rightarrow \frac{\partial y_\ell}{\partial w_\ell} = diag(\sigma'(w_\ell)) = diag(\sigma'(A_\ell y_{\ell-1}+b_\ell)) \Rightarrow \frac{\partial y_\ell}{\partial b_\ell} = diag(\sigma'(A_\ell y_{\ell-1}+b_\ell))$$

$$\frac{\partial y_\ell}{\partial y_{\ell-1}} = \frac{\partial y_\ell}{\partial w_\ell}\cdot\frac{\partial w_\ell}{\partial y_{\ell-1}} = diag(\sigma'(A_\ell y_{\ell-1}+b_\ell))\cdot A_\ell.$$

(b) $y_L = A_L y_{L-1} + b_L$

$\frac{\partial y_L}{\partial A_L} \in \mathbb{R}^{n_{L-1}\times 1}$, and $\left(\frac{\partial y_L}{\partial A_L}\right)_i = \frac{\partial y_L}{\partial (A_L)_i} = (y_{L-1})_i$.

$\Rightarrow \frac{\partial y_L}{\partial A_L} = y_{L-1}^T \in \mathbb{R}^{1\times n_{L-1}}$

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

Since $y_\ell = \sigma(A_\ell y_{\ell-1}+b_\ell) = \sigma(w_\ell) \in \mathbb{R}^{n_\ell}$

$$\frac{\partial y_\ell}{\partial (A_\ell)_{ij}} = \frac{\partial y_\ell}{\partial w_\ell}\cdot\frac{\partial w_\ell}{\partial (A_\ell)_{ij}} = diag(\sigma'(A_\ell y_{\ell-1}+b_\ell))\cdot\frac{\partial w_\ell}{\partial (A_\ell)_{ij}}$$

$$\left[\frac{\partial w_\ell}{\partial (A_\ell)_{ij}}\right]_k = \frac{\partial (w_\ell)_k}{\partial (A_\ell)_{ij}} = \frac{\partial}{\partial (A_\ell)_{ij}}\left[\sum_{m=1}^{n_{\ell-1}} (A_\ell)_{km}(y_{\ell-1})_m + b_\ell\right] = \delta_{ik}\cdot(y_{\ell-1})_j \in \mathbb{R}.$$

$$\Rightarrow \frac{\partial w_\ell}{\partial (A_\ell)_{ij}} \in \mathbb{R}^{n_\ell\times 1} \Rightarrow \frac{\partial w_\ell}{\partial (A_\ell)_{ij}} = (y_{\ell-1})_j\cdot e_i \qquad \left[e_i = \begin{pmatrix}0\\ 0\\ \vdots\\ 1\\ \vdots\\ 0\\ 0\end{pmatrix} \leftarrow \text{ith row} \in \mathbb{R}^{n_\ell}.\right]$$

$$\Rightarrow \frac{\partial y_L}{\partial (A_\ell)_{ij}} = \frac{\partial y_L}{\partial y_\ell}\cdot diag(\sigma'(A_\ell y_{\ell-1}+b_\ell))\cdot(y_{\ell-1})_j\cdot e_i$$

$$= (y_{\ell-1})_j\cdot\sigma'([A_\ell y_{\ell-1}+b_\ell]_i)\cdot\frac{\partial y_L}{\partial y_\ell}\cdot e_i.$$

$$= \sigma'([A_\ell y_{\ell-1}+b_\ell]_i)\cdot\left[\frac{\partial y_L}{\partial y_\ell}\right]_i\cdot(y_{\ell-1})_j.$$

$$B_\ell := \left(\frac{\partial y_L}{\partial y_\ell}\right)^T\cdot y_{\ell-1}^T \in \mathbb{R}^{n_\ell\times n_\ell}.$$

$$[B_\ell]_{ij} = \left[\frac{\partial y_L}{\partial y_\ell}\right]_i \cdot (y_{\ell-1})_j$$

$\in \mathbb{R}^{n_\ell \times n_L}$

$$C_\ell := \operatorname{diag}\left(\sigma'(A_\ell y_{\ell-1} + b_\ell)\right) \Rightarrow [C_\ell]_{ii} = \sigma'([A_\ell y_{\ell-1} + b_\ell]_i)$$

$$\frac{\partial y_L}{\partial (A_\ell)_{ij}} = (C_\ell)_{ii} \cdot (B_\ell)_{ij}$$

$$\Rightarrow [C_\ell B_\ell]_{ij} = \sum_{k=1}^{n_\ell} (C_\ell)_{ik} (B_\ell)_{kj} = (C_\ell)_{ii} (B_\ell)_{ij} = \frac{\partial y_L}{\partial (A_\ell)_{ij}}$$

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

$\blacksquare$