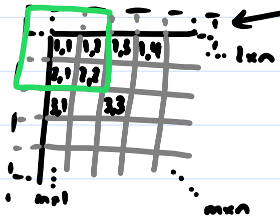
MFDNN

Homework 4

24/03/26

1\. 

← zero-padding all round image assumed (Problem didn't define $X_{0,:} = 0$ specifically)

Desired:

$$\sum\left(\begin{pmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{pmatrix} \odot \begin{pmatrix} 0 & 0 & 0 \\ 0 & x_{11} & x_{12} \\ 0 & x_{21} & x_{22} \end{pmatrix}\right) = X_{21} - X_{11}$$

Suggests $w_8 = 1$, $w_5 = -1$, $w_i = 0 \ \forall$ other i

$$\sum\left(\begin{pmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{pmatrix} \odot \begin{pmatrix} 0 & 0 & 0 \\ 0 & x_{11} & x_{12} \\ 0 & x_{21} & x_{22} \end{pmatrix}\right) = X_{12} - X_{11}$$

$w_6 = 1$, $w_5 = -1$, $w_i = 0 \ \forall$ other i

So for $\omega \in \mathbb{R}^{2\times3\times3}$:

$$\omega = \left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix}\right)$$

Sanity check:

$$Y_{\ell,i,j} = \sum_{\alpha=1}^{3}\sum_{\beta=1}^{3} \omega_{\ell,\alpha,\beta} X'_{i+\alpha-1,\, j+\beta-1}$$

← padded X

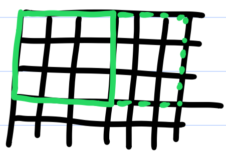
$$Y_{1,1,1} = \omega_{22} X_{11} + \omega_{32} X_{21} = -X_{11} + X_{21} \ ✓$$

$$Y_{2,1,1} = \omega_{22} X_{11} + \omega_{23} X_{12} = -X_{11} + X_{12} \ ✓$$

$$Y_{1,2,2} = \omega_{22} X_{22} + \omega_{32} X_{32} = -X_{22} + X_{32} \ ✓$$

2\. The AvgPool2d operation conceptually simply outputs the average of the 'window' being considered. Since a convolution outputs the sum of the product of the kernel weights with the window elements simply define $\omega = \frac{1}{k^2}\underline{1} \in \mathbb{R}^{k\times k}$ and use stride of k to ensure no overlaps of kernel.

↳ or simply take ω as a constant multiplier/kernel



3\. $\omega = \left((0.299), (0.587), (0.114)\right)$

$\mathbb{R}^{1\times1}$

4. Consider a block, $B \in \mathbb{R}^{p\times q}$ $(p \le m, q \le n)$ in $X \in \mathbb{R}^{m\times n}$ where p and q are the kernel size of the max pool.
Let X_{max} be (one of) the largest element of B. $\rho(B) = X_{max}$ by definition of X_{max}.
Now we apply the activation function to this block's output to give, $\sigma(\rho(B)) = \sigma(X_{max})$.

Now, $X_{max} \geq X_{ij}$ for any $X_{ij} \in B$ by definition of X_{max}.
Since σ is non-decreasing, $X_{max} \geq X_{ij} \Rightarrow \sigma(X_{max}) \geq \sigma(X_{ij})$ for all $X_{ij} \in B$.
So if we were to apply the max pool operation now on $\sigma(B)$, clearly, $\rho(\sigma(B)) = \sigma(X_{max})$.
since $\sigma(X_{max})$ is also the largest element of $\sigma(B)$ (by above).

Since this argument is true for any and all blocks $B \in X$, this implies $\rho(\sigma(X)) = \sigma(\rho(X))$.
and hence all elements of output $\mathbb{R}^{k\times l}$ matrix

5. (Workings)

$$f_i(\underline{a},\underline{b}) := \| \underline{P}(Y_i) - \underline{f}_{\underline{a},\underline{b}}(\underline{X}_i) \|^2$$

$$z := \underline{a}^T \underline{x} + b$$

$$= \| \underline{P}(Y_i) - \{ \sigma(-z), \sigma(z) \} \|^2$$

$$= \underbrace{(1-Y_i)\tfrac{1}{2}}_{\substack{\text{if } Y_i=-1,\ =1 \\ \text{if } Y_i=1,\ =0}} \| \{1,0\} - \{\sigma(-z),\sigma(z)\} \|^2 + \underbrace{(1+Y_i)\tfrac{1}{2}}_{\substack{\text{if } Y_i=1,\ =1 \\ \text{if } Y_i=-1,\ =0}} \| \{0,1\} - \{\sigma(-z),\sigma(z)\} \|^2$$

$$= \tfrac{1}{2}(1-Y_i)\left((1-\sigma(-z))^2 + \sigma(z)^2 \right) + \tfrac{1}{2}(1+Y_i)\left(\sigma(-z)^2 + (1-\sigma(z))^2 \right)$$

$$=: \ell(z, Y_i) = \ell(\underline{a}^T \underline{x} + b, Y_i)$$

6.(a) $$\frac{\partial y_L}{\partial b_L} = \frac{\partial}{\partial b_L}(A_L y_{L-1} + b_L) = \boxed{1}$$

$$\frac{\partial y_L}{\partial y_{L-1}} = \frac{\partial}{\partial y_{L-1}}(A_L y_{L-1} + b_L) = \boxed{A_L}$$

For $\ell = 1, \dots, L-1$,

$$\frac{\partial \underline{y}_\ell}{\partial \underline{b}_\ell} = \begin{pmatrix} \frac{\partial y_{\ell 1}}{\partial b_{\ell 1}} & \frac{\partial y_{\ell 1}}{\partial b_{\ell 2}} & \cdots & \frac{\partial y_{\ell 1}}{\partial b_{\ell n_\ell}} \\ \frac{\partial y_{\ell 2}}{\partial b_{\ell 1}} & \frac{\partial y_{\ell 2}}{\partial b_{\ell 2}} & \cdots & \frac{\partial y_{\ell 2}}{\partial b_{\ell n_\ell}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_{\ell n_\ell}}{\partial b_{\ell 1}} & \frac{\partial y_{\ell n_\ell}}{\partial b_{\ell 2}} & \cdots & \frac{\partial y_{\ell n_\ell}}{\partial b_{\ell n_\ell}} \end{pmatrix}$$

Define $\tilde{\underline{y}}_\ell = A_\ell y_{\ell-1} + \underline{b}_\ell$ (i.e. y_ℓ pre-activation)

$$\frac{\partial \tilde{y}_{\ell i}}{\partial b_{\ell j}} = \frac{\partial}{\partial b_{\ell j}}\left([A_\ell y_{\ell-1} + \underline{b}_\ell]_i\right)$$
$$= \frac{\partial}{\partial b_{\ell j}}\left([A_\ell y_{\ell-1}]_i + b_{\ell i}\right)$$
$$= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

$$\frac{\partial y_{\ell i}}{\partial b_{\ell j}} = \frac{\partial}{\partial b_{\ell j}}\left([\underline{\sigma}(\tilde{\underline{y}}_\ell)]_i\right)$$
$$= \frac{\partial \tilde{y}_{\ell i}}{\partial b_{\ell j}}[\underline{\sigma}'(\tilde{\underline{y}}_\ell)]_i = \frac{\partial \tilde{y}_{\ell i}}{\partial b_{\ell j}} \sigma'(\tilde{y}_{\ell i})$$

by chain rule and elementwise application

$$= \begin{cases} 0 & \text{if } i \neq j \\ \sigma'(\tilde{y}_{\ell i}) & \text{if } i = j \end{cases} \text{ by above}$$

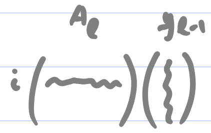
Hence, only diagonal entries of $\frac{\partial \underline{y}_\ell}{\partial \underline{b}_\ell}$ are non-zero (i=j)

with $\frac{\partial y_{\ell i}}{\partial b_{\ell i}} = \sigma'(\tilde{y}_{\ell i}) = \sigma'([A_\ell y_{\ell-1} + b_\ell]_i)$

That is, $\boxed{\frac{\partial \underline{y}_\ell}{\partial \underline{b}_\ell} = \mathrm{diag}(\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell))}$ ∎

For $\ell = 2, \dots, L-1$,

For $\frac{\partial \underline{y}_\ell}{\partial y_{\ell-1}}$, consider entries once more:

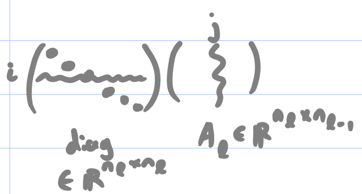


$$\frac{\partial \tilde{y}_{\ell i}}{\partial (y_{\ell-1})_j} = \frac{\partial}{\partial (y_{\ell-1})_j}\left([A_\ell y_{\ell-1}]_i + b_{\ell i}\right)$$
$$= \frac{\partial}{\partial (y_{\ell-1})_j}\left(\sum_{k=1}^{n_{\ell-1}} (A_\ell)_{ik}(y_{\ell-1})_k + b_{\ell i}\right)$$
$$= (A_\ell)_{ij}$$

So $$\frac{\partial (y_\ell)_i}{\partial (y_{\ell-1})_j} = \frac{\partial}{\partial (y_{\ell-1})_j}\left(\sigma((\tilde{y}_\ell)_i)\right) = \frac{\partial (\tilde{y}_\ell)_i}{\partial (y_{\ell-1})_j}\sigma'(\tilde{y}_{\ell i}) = (A_\ell)_{ij}\sigma'((\tilde{y}_\ell)_i)$$
$$= \sigma'((\tilde{y}_\ell)_i)(A_\ell)_{ij}$$
$$= [\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell)]_i (A_\ell)_{ij}$$

$\updownarrow$ elements equivalent

$$[\mathrm{diag}(\underline{\sigma}'(A_\ell y_{\ell-1} + b_\ell))A_\ell]_{ij} = \sum_{k=1}^{n_\ell} \underbrace{[\mathrm{diag}(\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell))]_{ik}}_{\neq 0 \text{ iff } k=i}(A_\ell)_{kj} = [\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell)]_i (A_\ell)_{ij}$$



$$\Rightarrow \boxed{\frac{\partial \underline{y}_\ell}{\partial y_{\ell-1}} = \mathrm{diag}(\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell))A_\ell}$$

as given ∎

6.(b) $$\left(\frac{\partial y_L}{\partial A_L}\right)_{ij} \equiv \frac{\partial y_L}{\partial (A_L)_{ij}} = \frac{\partial}{\partial (A_L)_{1j}}\left(A_L \underline{y}_{L-1} + b_L\right)$$

$A_L \in \mathbb{R}^{1 \times n_{\ell-1}}$
so only $i=1$ defined

$$= \frac{\partial}{\partial (A_L)_{1j}}\left(\sum_{k=1}^{n_{\ell-1}} (A_L)_{1,k} (y_{L-1})_k\right)$$

$$= (y_{L-1})_j \quad \Rightarrow \quad \frac{\partial y_L}{\partial A_L} = \left((y_{L-1})_1, (y_{L-1})_2, \dots, (y_{L-1})_{n_{\ell-1}}\right)$$

$$= y_{L-1}^T \quad \blacksquare$$

For $\ell = 1, \dots, L-1$,

$$\frac{\partial y_L}{\partial (A_\ell)_{ij}} = \frac{\partial y_L}{\partial y_\ell} \frac{\partial y_\ell}{\partial (A_\ell)_{ij}} = \left(\frac{\partial y_L}{\partial y_{L-1}} \frac{\partial y_{L-1}}{\partial y_{L-2}} \cdots \frac{\partial y_{\ell+1}}{\partial y_\ell}\right) \frac{\partial y_\ell}{\partial (A_\ell)_{ij}}$$ by chain rule

$$= (A_L)\left(\mathrm{diag}\left(\underline{\sigma}'(A_L y_{L-1} + \underline{b}_L)\right) A_L\right)\left(\mathrm{diag}\left(\underline{\sigma}'(A_{L-1} y_{L-2} + \underline{b}_{L-1})\right) A_{L-1}\right) \cdots \left(\mathrm{diag}\left(\underline{\sigma}'(A_{\ell+1} \underline{y}_\ell + \underline{b}_{\ell+1})\right) A_{\ell+1}\right) \frac{\partial y_\ell}{\partial (A_\ell)_{ij}}$$
by part (a)

$$\frac{\partial y_\ell}{\partial (A_\ell)_{ij}} = \frac{\partial}{\partial (A_\ell)_{ij}}\left(\underline{\sigma}\left[\sum_{p=1}^{n_\ell}\left(\underline{e}_p \sum_{k=1}^{n_{\ell-1}} (A_\ell)_{pk} (\underline{y}_{\ell-1})_k\right) + \underline{b}_\ell\right]\right)$$

$$= \underline{e}_i (y_{\ell-1})_j \underline{\sigma}'(\tilde{\underline{y}}_\ell)$$