MFDNN

Homework 7

24/04/26

1.(a) $\lim_{\beta\to\infty}\left(\frac{1}{\beta}\log\left(\sum_{i=1}^{n}\exp(\beta x_i)\right)\right) = \lim_{\beta\to\infty}\left(\frac{1}{\beta}\left(\log(\exp(\beta x_j)) + \log\left(1+\sum_{\substack{i=1,\\ i\neq j}}^{n}\frac{\exp(\beta x_i)}{\exp(\beta x_j)}\right)\right)\right)$

since $\log(a+b)=\log(a)+\log(1+\frac{b}{a})$; where $x_j := \max\{x_1,\dots,x_n\}$

$$= \lim_{\beta\to\infty}\left(x_j + \frac{1}{\beta}\log\left(1+\sum_{\substack{i=1,\\ i\neq j}}^{n}\exp(\beta(x_i-x_j))\right)\right)$$

Since x_j is the max x, $x_i - x_j \leq 0 \ \forall i$

$\Rightarrow$ argument of exponential is negative

So in limit $\beta\to\infty$, $\exp(\beta(x_i-x_j))\to 0 \ \forall i$

(if the maximum is not unique, then the summation is a finite integer $<n$)
(each shared max adds 1)

In any case,

$$= \lim_{\beta\to\infty}\left(x_j + \frac{\log(1+\gamma)}{\beta}\right), \ \gamma\to 0 \text{ if max unique, else } 0<\gamma<n$$

$$= x_j + 0$$

$= \max\{x_1,\dots,x_n\}$ as required ∎

(b) $\underline{\nabla}\gamma_1 = \left(\frac{\partial}{\partial x_1}\gamma_1, \frac{\partial}{\partial x_2}\gamma_1, \dots, \frac{\partial}{\partial x_n}\gamma_1\right)$

$\frac{\partial}{\partial x_j}\gamma_1 = \frac{\partial}{\partial x_j}\left(\log\left(\sum_{i=1}^{n}e^{x_i}\right)\right)$

$= e^{x_j}\cdot\frac{1}{\sum_{i=1}^{n}e^{x_i}}$

$$= \left(\frac{e^{x_1}}{\sum_{i=1}^{n}e^{x_i}}, \frac{e^{x_2}}{\sum_{i=1}^{n}e^{x_i}}, \dots, \frac{e^{x_n}}{\sum_{i=1}^{n}e^{x_i}}\right)$$

Clearly, $(\underline{\nabla}\gamma_1(\underline{x}))_k = \frac{\exp(x_k)}{\sum_{i=1}^{n}\exp(x_i)} = \mu_k(\underline{x})$

True $\forall k \Rightarrow \underline{\nabla}\gamma_1 = \mu$ ∎

(c) $\frac{\partial}{\partial x_j}\gamma_\beta = \frac{1}{\beta}\cdot\frac{\beta\exp(\beta x_j)}{\sum_{i=1}^{n}\exp(\beta x_i)}$

$$\underline{\nabla}\gamma_\beta(\underline{x}) = \sum_{j=1}^{n}\frac{\exp(\beta x_j)}{\sum_{i=1}^{n}\exp(\beta x_i)}\underline{e}_j = \sum_{j=1}^{n}\frac{1}{\sum_{i=1}^{n}\exp(\beta(x_i-x_j))}\underline{e}_j$$

- For term where $j=i_{max}$, $x_i - x_j < 0 \Rightarrow \exp(\beta(x_i-x_j))\to 0$ as $\beta\to\infty$
  for all $i\neq j$

  For $i=j$, $x_i - x_j = 0 \Rightarrow \exp(\beta(x_i-x_j)) = 1$

  So for this term, denominator is 1 and so $\underline{e}_j = \underline{e}_{i_{max}}$ term is kept.
- For all other j, $\exists i$ s.t. $x_i > x_j \Rightarrow x_i - x_j > 0 \Rightarrow \exists i$ for which $\exp(\beta(x_i-x_j))\to\infty$ as $\beta\to\infty$

  So the denominator tends to infinity and the term vanishes and $\underline{e}_j$ is lost.

So as $\beta\to\infty$, $\underline{\nabla}\gamma_\beta(\underline{x})\to\underline{e}_{i_{max}}$ ∎ (i.e. only $j=i_{max}$ term ($\underline{e}_{i_{max}}$) is preserved) including additional zero values due to padding

2. Num convolution additions = Num convolution multiplications (AW 5.5)

$= h_{out}^2 \times k_h^2 \times C_{in} \times C_{out}$

(h_{out} = output image width/height $= \left\lfloor \frac{h_{in} - k_h + 2P}{S} + 1 \right\rfloor$) applicable to pooling too

Num linear layer multiplications $= C_{out} C_{in}$

Num linear layer additions $= C_{out} (\underbrace{C_{in} - 1}_{\text{summation}} + \underbrace{1}_{\text{bias}})$

$= C_{out} C_{in}$

$Y_i = (Ax)_i + b_i \quad i = 1, \dots, C_{out}$

$= \sum_{j=1}^{C_{in}} A_{ij} x_j + b_i$

So we have:

convolution num	C_{in}	C_{out}	k_h	(h_{in}	P	S)	h_{out}
1	3	64	11	227	0	4	55
2	64	192	5	27	2	1	27
3	192	384	3	13	1	1	13
4	384	256	3	13	1	1	13
5	256	256	3	13	1	1	13 no pooling

linear num	C_{in} ($256 \times 6 \times 6$)	C_{out}
1	9216	4096
2	4096	4096
3	4096	1000 ← num classes

$\Rightarrow$ Convolutional layers total $= \underbrace{2}_{\text{add \& mult}} \times (\text{conv1} + \text{conv2} + \text{conv3} + \text{conv4} + \text{conv5})$ — each $C_{in} \times C_{out} \times k_h^2 \times h_{out}^2$ from table

$= 1{,}311{,}133{,}056$

Linear layers total $= 2 \times (\text{lin1} + \text{lin2} + \text{lin3})$ — each $C_{in} \times C_{out}$

$= 117{,}243{,}904$

3. (Workings)

$$\omega \in \mathbb{R}^{\underbrace{C_{out}}_{\text{normalised}} \times C_{in} \times f_1 \times f_2}$$

$$\underbrace{Y_{k,\ell,i,j}}_{\text{out channel}} = \sum_{\gamma=1}^{C_{in}} \sum_{\alpha=1}^{f_1} \sum_{\beta=1}^{f_2} \omega_{\ell,\gamma,\alpha,\beta} X_{k,\gamma,i+\alpha-2,j+\beta-2} + b_\ell$$

$$BN_{\gamma,\beta}(Y)[b,:,i,j] = \gamma[:] \frac{Y[b,:,i,j] - \hat{\mu}[:]}{\sqrt{\hat{\sigma}^2[:] + \varepsilon}} + \beta[:]$$

$$= \frac{\gamma[:] Y[b,:,i,j]}{\sqrt{\hat{\sigma}^2[:] + \varepsilon}} \overbrace{- \frac{\gamma[:]\hat{\mu}[:]}{\sqrt{\hat{\sigma}^2[:] + \varepsilon}} + \beta[:]}^{\text{new bias}}$$

$$Y'_{k,\ell,i,j} = \frac{\gamma[\ell]}{\sqrt{\hat{\sigma}^2[\ell] + \varepsilon}} \left(\sum\sum\sum \omega X + b_\ell \right)$$

$$= \overbrace{\frac{\gamma[\ell]}{\sqrt{\hat{\sigma}^2[\ell] + \varepsilon}}}^{\text{weight multiplier}} \underbrace{\sum\sum\sum \omega X}_{\text{use to define } \omega'} + \overbrace{\frac{\gamma[\ell]}{\sqrt{\hat{\sigma}^2[\ell] + \varepsilon}} b_\ell}^{\text{to be inc. in bias}}$$

$$A \in \mathbb{R}^{\underbrace{C_{out}}_{\text{normalised}} \times C_{in}}$$

precision? question

4.(a) $$\frac{\partial y_L}{\partial y_{L-1}} = \frac{\partial}{\partial y_{L-1}}\left(A_{w_L} y_{L-1} + b_L \mathbb{1}_{n_L}\right)$$

$$= A_{w_L}$$

$$\left(\frac{\partial y_\ell}{\partial y_{\ell-1}}\right)_{ij} = \frac{\partial}{\partial (y_{\ell-1})_j}\left(\sigma\left(\sum_k (A_{w_\ell})_{ik}(y_{\ell-1})_k + b_\ell\right)\right)$$

$$= (A_{w_\ell})_{ij}\cdot\sigma_i'\left(A_{w_\ell} y_{\ell-1} + b_\ell \mathbb{1}_{n_\ell}\right)$$

$$\Rightarrow \frac{\partial y_\ell}{\partial y_{\ell-1}} = \mathrm{diag}\left(\sigma'\left(A_{w_\ell} y_{\ell-1} + b_\ell \mathbb{1}_{n_\ell}\right)\right) A_{w_\ell} \quad \blacksquare$$

(see HW4.6(a) equivalence workings)

$$\left(\frac{\partial y_L}{\partial w_\ell}\right)_{1,i} = \sum_{k=1}^{n_\ell}\left(\frac{\partial y_L}{\partial y_\ell}\right)_{1,k}\left(\frac{\partial y_\ell}{\partial w_\ell}\right)_{k,i} \qquad i=1,\dots,f_\ell$$

$$= \sum_{k=1}^{n_\ell}\left(\frac{\partial y_L}{\partial y_\ell}\right)_{1,k}\sigma_k'\left(A_{w_\ell} y_{\ell-1} + b_\ell \mathbb{1}_{n_\ell}\right)(y_{\ell-1})_{k+i-1}$$

$$= \sum_{k=1}^{n_{\ell-1}-f_\ell+1}\left(\frac{\partial y_L}{\partial y_\ell}\right)_{1,k}\sigma_k'\left(A_{w_\ell} y_{\ell-1} + b_\ell \mathbb{1}_{n_\ell}\right)(y_{\ell-1})_{k+i-1}$$

... ?

$$= \sum_{m=1}^{n_{\ell-1}}(C_{v_\ell^T})_{im}(y_{\ell-1})_m = (C_{v_\ell^T} y_{\ell-1})_i$$

$$= (C_{v_\ell^T} y_{\ell-1})^T_{1,i}$$

$\frac{\partial y_L}{\partial y_\ell} \in \mathbb{R}^{1\times n_\ell}$

$\mathrm{diag}(\sigma'(\dots)) \in \mathbb{R}^{n_\ell\times n_\ell}$

$y_{\ell-1} \in \mathbb{R}^{n_{\ell-1}\times 1}$

$v_\ell \in \mathbb{R}^{1\times n_\ell}$

$v_\ell^T \in \mathbb{R}^{n_\ell\times 1}$

$C_{v_\ell^T} \in \mathbb{R}^{f_\ell\times n_{\ell-1}}$

$\frac{\partial y_L}{\partial w_\ell} \in \mathbb{R}^{1\times f_\ell} \ni (C_{v_\ell^T} y_{\ell-1})^T$

$C_{v_\ell^T} y_{\ell-1} \in \mathbb{R}^{f_\ell\times 1}$

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

$$= \delta_{k,i+j-1}\,(y_{\ell-1})_k\,\sigma_i'\left(A_{w_\ell} y_{\ell-1} + b_\ell \mathbb{1}_{n_\ell}\right)$$

$$= (y_{\ell-1})_{i+j-1}\,\sigma_i'\left(A_{w_\ell} y_{\ell-1} + b_\ell \mathbb{1}_{n_\ell}\right)$$

$$\frac{\partial (A_{w_\ell})_{ik}}{\partial (w_\ell)_j} = \delta_{k,i+j-1}$$

$(w_\ell)_j$ located in positions in A_{w_ℓ} $\leftarrow$

$(1,j)$
$(2,j+1)$
$\vdots$
$(i,j+(i-1))$

$$A_{w_\ell} = \begin{pmatrix} (w_\ell)_1 & \cdots & (w_\ell)_{f_\ell} & 0 & & 0\cdots 0 \\ 0 & (w_\ell)_1 & \cdots & (w_\ell)_{f_\ell} & & 0\cdots 0 \\ 0 & & \ddots & & \ddots & \vdots \\ \vdots & & & & & 0 \\ 0 & & 0 & (w_\ell)_1 & \cdots & (w_\ell)_{f_\ell} \end{pmatrix}$$

(as in HW1.5)
(or Chp3 Slide 19)

(a) (cont.) $\frac{\partial y_L}{\partial b_\ell} = \frac{\partial y_L}{\partial y_\ell}\frac{\partial y_\ell}{\partial b_\ell}$

$$\left(\frac{\partial y_\ell}{\partial b_\ell}\right)_i = \frac{\partial}{\partial b_\ell}\left(\sigma\left(\sum_k^{n_\ell}(A_{w_\ell})_{ik}(y_{\ell-1})_k + b_\ell\right)\right)$$
$$= \sigma_i'(A_{w_\ell}y_{\ell-1} + \underline{b}_\ell \mathbb{1}_{n_\ell})$$

$$= \frac{\partial y_L}{\partial y_\ell}\,\mathrm{diag}(\underline{\sigma}'(A_{w_\ell}y_{\ell-1} + \underline{b}_\ell \mathbb{1}_{n_\ell}))\,\mathbb{1}_{n_\ell}$$

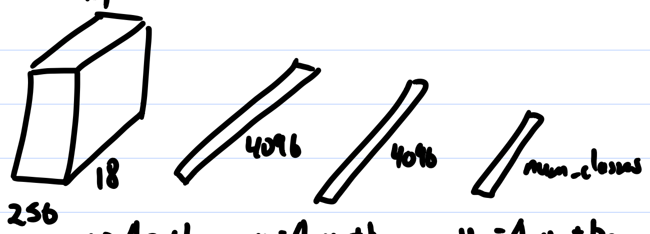
$$= \underline{v}_\ell \mathbb{1}_{n_\ell} \quad \square$$

(b) In the forward pass, matrix-vector products with A_{w_i} are used to perform convolutions (for $A_{w_i}\underline{y}_{i-1}$). In backpropagation, calculations will require updating $\frac{\partial y_L}{\partial y_{\ell-1}}$ using right multiplication by A_{w_ℓ}, which should use $A_{w_\ell}^T$ in transpose-convolutions.

Calculations will also require $\frac{\partial y_L}{\partial w_\ell}$ which uses the convolutional (i.e. matrix) operator $C_{v_\ell^T}$. This should be performed using convolution since part of a matrix-vector product. Note that computing $C_{v_\ell^T}$ (by extension $\underline{v}_\ell$) will req. transpose convolutions (for $\frac{\partial y_L}{\partial y_\ell}$) and regular convolutions (for $A_{w_\ell}y_{\ell-1}$ in $\underline{\sigma}'$ function).

5. (workings)

(a) Net 1



$y_1 = A_1 x + b_1$ $\quad y_2 = A_2 y_1 + b_2$ $\quad y_3 = A_3 y_2 + b_3$

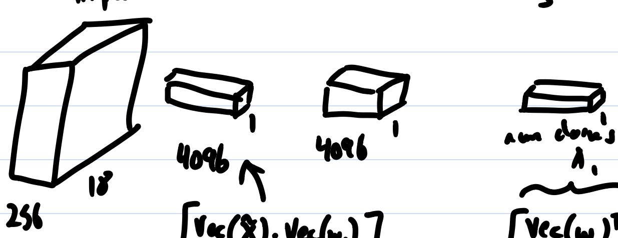
output

$A_1 \in \mathbb{R}^{4096 \times (256 \cdot 18 \cdot 18)}$

$A_2 \in \mathbb{R}^{4096 \times 4096}$

$A_3 \in \mathbb{R}^{num\_classes \times 4096}$

Net 2



$w_i \in \mathbb{R}^{256 \times 18 \times 18}$

convolution: $\begin{bmatrix} Vec(\tilde{X}) \cdot Vec(w_1) \\ \vdots \\ Vec(\tilde{X}) \cdot Vec(w_c) \end{bmatrix} + b_1 = \underbrace{\begin{bmatrix} Vec(w_1)^T \\ \vdots \\ Vec(w_c)^T \end{bmatrix}}_{A_1} vec(\tilde{X}) + b_1$

$c = 4096$