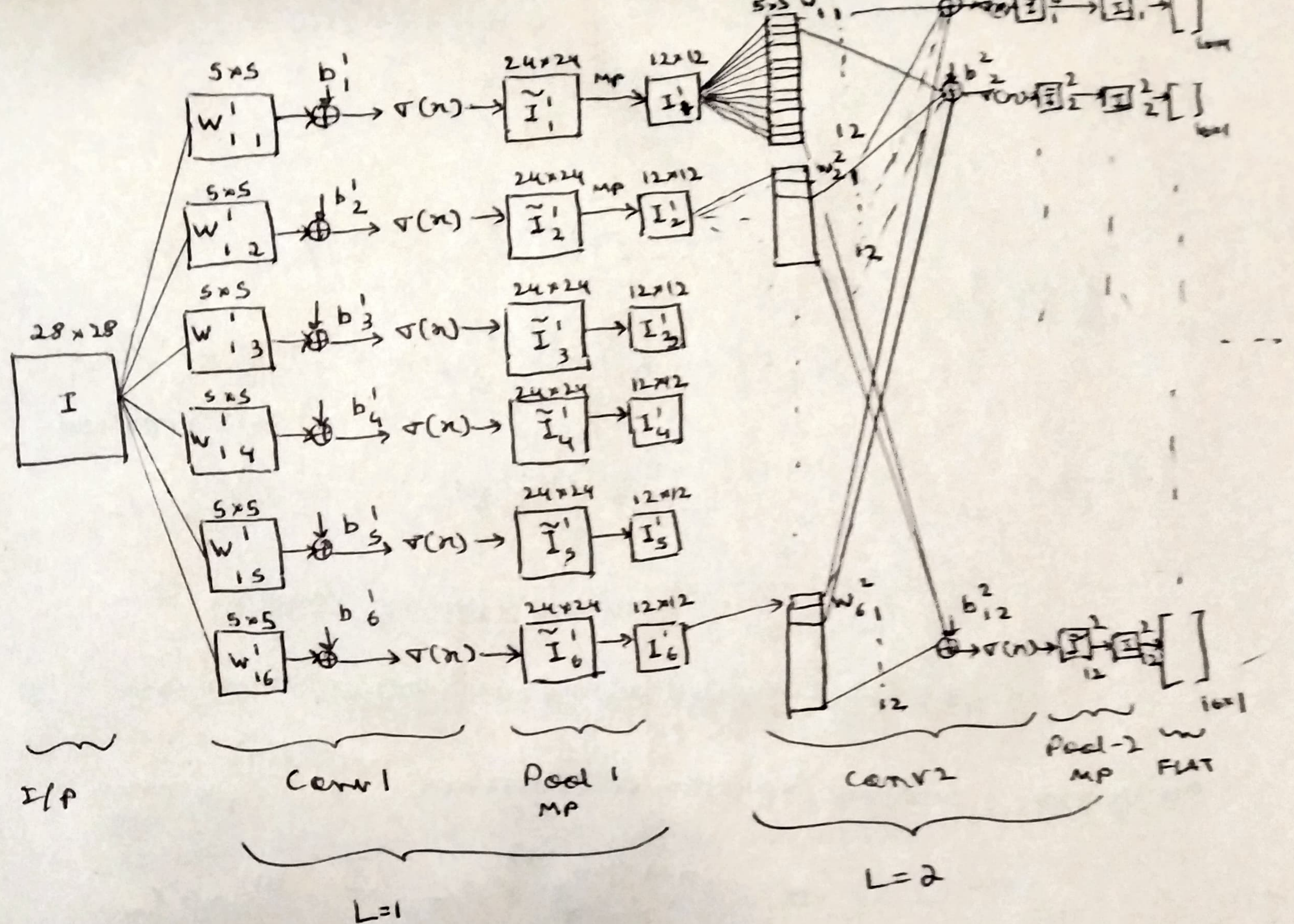
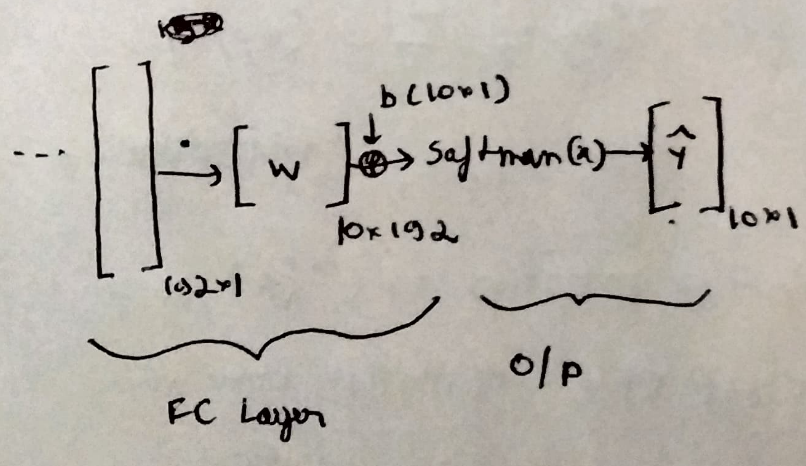
HW-3 ECE 792(061) SOMSHUBHRA ROY





$$L(y, \hat{y}) = \sum_{i=1}^{10} y_i \log(\hat{y}_i)$$

1> <u>Softmax layer</u>

$$\frac{\delta L}{\delta \hat{y}_i} = -y_i/\hat{y}_i \quad i \in [1, 10]$$

2> <u>output of Dense layer</u>

$$\delta L/\delta z_j = \hat{y}_j - y_j \quad j \in [1, 10]$$

3> <u>Dense layer</u>

$$\delta L/\delta w_{jk} = \delta L/\delta z_j \times x_k \qquad x_k = k^{th} \text{ input of dense layer}$$

$$\delta L/\delta b_j = \delta L/\delta z_j$$

4> <u>Maxpool 2</u>

$$\delta L/\delta \tilde{x}^2_i = \delta L/\delta z_j \times 1 \text{ if } x_i = \text{max in pooling window}$$
$$= 0 \text{, otherwise}$$

5) ~~Conv 2 sigmoid 2~~ Conv 2 sigmoid 2

where, $\sigma'(x) = \sigma(x)(1-\sigma(x))$

~~$\delta L/\delta x$~~ :

$$\delta L/\delta y^2_{jk} = \sum_i \sum_{p,q} \delta L/\delta \hat{x}^2_{ipq} \times w^2_{jpq} \times \sigma'(y^2_{jk})$$

when $\hat{x}^2_{ipq}$ is output of maxpool 1 and corresponds to ~~max index~~

max value at position (i, p, q) of the input of maxpool 1

& $\sigma'(y^2_{jk})$ is evaluated at output of Conv 2 sigmoid 2 layer,

~~6) Conv bias~~

$$\delta L/\delta b^2_j = \sum_k \delta L/\delta y^2_{jk}$$ bias gradient

$$\delta L/\delta w^2_{jpq} = \sum_k \delta L/\delta y^2_{jk} \; \hat{x}^1_{i+p,\, q+k}$$

where $\hat{x}^1_{(i+p,\, q+k)}$ is ~~corresponds to~~ output of maxpool 1 corresponding

to window containing element at position $(i+p, q+k)$

4) Layer 1 (conv 1 sigmoid 1 Maxpool 1)

$$\delta L/\delta y^1_{pq} = \sum_j \sum_k \delta L/\delta y^2_{jk} \times w^2_{jp(k+u)(q+v)} \times \sigma'(y^1_{pq})$$

~~for~~ following above rules.

$$\therefore \delta L/\delta w = \delta L/\delta \hat{y} \times \delta \hat{y}/\delta w = (\hat{y} - y) \times (x)^T$$

$$\delta L/\delta b = \delta L/\delta \hat{y} \times \delta \hat{y}/\delta b = (\hat{y} - y)$$

$$\delta L/\delta b^l_p = \sum_q \sum_i \sum_j \left(\frac{\delta L}{\delta z^i_{jq}} \; \frac{\delta z^i_{jq}}{\delta b^1_p}\right)$$

$$\delta L/\delta w^l_{pq} = \sum_i \sum_j \left(\frac{\delta L}{\delta z^i_{jq}} \; \frac{\delta z^i_{jq}}{\delta w^1_{pq}}\right)$$

[Ans:]

3) Given

a) 2 K-D gaussian distributions $N(\mu_0, \Sigma_0), N(\mu_1, \Sigma_1)$

R.T.P

$$D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T \Sigma^{-1} (\mu_1-\mu_0) - K + \log\left(\frac{det(\Sigma_1)}{det(\Sigma_0)}\right)\right)$$

Proof: We know that the KL divergence between 2 PDFs are given

as $$D_{KL}(P_1 \| P_2) = \sum_{x \in X} P_1(x) \log\left(\frac{P_1(x)}{P_2(x)}\right)$$

$$= -\sum_{x \in X} P_1(x) \log\left(\frac{P_2(x)}{P_1(x)}\right)$$

$$= +E_{P_1} \log\left(\frac{P_1}{P_2}\right)$$

pdf of a multivariate Gaussian distribution is given as

$$p(x) = \frac{1}{(2\pi)^{N/2} det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

In this case $D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = E_{N_0}(\log N_0 - \log N_1)$

$$= \frac{1}{2} E_{N_0}\left[-\log(det(\Sigma_0)) - (x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) + \log(det(\Sigma_1)) - (x-\mu_1)^T \Sigma_0^{-1}(x-\mu_1)\right]$$

$$= \frac{1}{2}\log\left(\frac{det(\Sigma_1)}{det(\Sigma_0)}\right) + \frac{1}{2} E_{N_0}\left[-(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) + (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right]$$

$$= \frac{1}{2}\log\left(\frac{det(\Sigma_1)}{det(\Sigma_0)}\right) + \frac{1}{2} E_{N_0}\left[-tr(\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^T) + tr(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T)\right]$$

$$= \frac{1}{2}\log\left(\frac{det(\Sigma_1)}{det(\Sigma_0)}\right) - \frac{1}{2}K + \frac{1}{2} tr(\Sigma_1^{-1}(\Sigma_0 + \mu_0\mu_0^T - 2\mu_1\mu_0^T + \mu_1\mu_1^T))$$

$$= \frac{1}{2}\left(\log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\right) - K + tr(\Sigma_1^{-1}\Sigma_0) + tr(\mu_0^T \Sigma_1^{-1} \mu_0 - 2\mu_0^T \Sigma_1^{-1}\mu_1 + \mu_1^T \Sigma_1^{-1}\mu_1)\right)$$

$$= \frac{1}{2}\left(\log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\right) - K + tr(\Sigma_1^{-1}\Sigma_0) - (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1 - 0)\right)$$

$$= \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - K + \log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\right)$$

[Proved]

b) KL-Divergence is not a metric loss

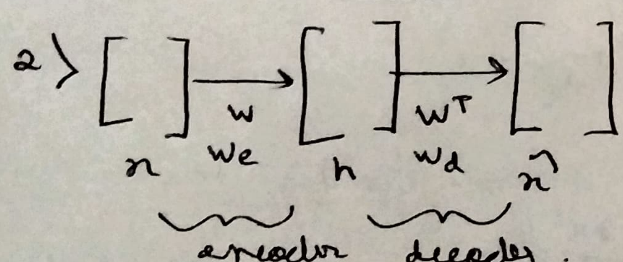
∴ it does not follow the symmetry property of metric space

$$\therefore D_{KL}(N(\mu_0,\Sigma_0)\,\|\,N(\mu_1,\Sigma_1)) \neq D_{KL}(N(\mu_1,\Sigma_1)\,\|\,N(\mu_0,\Sigma_0))$$

[Proved]

---

2)



$$L = (x-\hat{x})^2 \times \frac{1}{2N}$$

$$= (x - W_d W_e x - W_d b - c)^2 \times \frac{1}{2N}$$

$$h = W_e x + b$$

$$\hat{x} = W_d h + c$$

$$= W_d W_e x + W_d b + c.$$

$$\frac{\delta L}{\delta W_d} = (0 - W_e x - b - 0) \times \frac{1}{N}(x - W_d W_e x - W_d b - c)$$

$$\frac{\delta L}{\delta W_e} = \frac{1}{N}(x - W_d W_e x - W_d b - c)(0 - (W_d x)^T - 0 - 0)$$

$\frac{\delta L}{\delta W}$ ~~wrt~~ $W_d = W^T$

$W_e = W$

$L = (x - W^T W x - W^T b - c)^2 * \frac{1}{2N}$

$\frac{\delta L}{\delta W} = (0 - (x + x^T) W - b)(x - W^T W x - W^T b - c)\frac{1}{N}$

$= (-(x + x^T) W - b) \odot L'(x, \hat{x})$

$\frac{\delta L}{\delta W_d} = (-Wx - b) L'(x, \hat{x})$

$\frac{\delta L}{\delta W_e} = (-(W^T x)^T) L'(x, \hat{x})$

$= (-(W x^T)) L'(x, \hat{x})$

$\therefore \frac{\delta L}{\delta W_d} + \frac{\delta L}{W_e} = \frac{\delta L}{\delta W}$ [Proved.