Batch normalisation : Derivation for alternative backward propagation proof.

let us define

$$X = \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{Bmatrix} = \begin{Bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \vdots & & & \\ x_{M,1} & x_{M,2} & \cdots & x_{M,D} \end{Bmatrix} \in [M \times D]$$

$$\mu = [\mu_1, \mu_2, \cdots \mu_D] = \left[\frac{1}{M}\sum_{k=1}^{M} x_{k,1}, \frac{1}{M}\sum_{k=1}^{M} x_{k,2} \cdots \frac{1}{M}\sum_{k=1}^{M} x_{k,D}\right] \in [1 \times D]$$

$$\sigma^2 = [\sigma_1, \sigma_2 \cdots \sigma_D] = \left[\frac{1}{M}\sum_{k=1}^{M}(x_{k,1}-\mu_1)^2, \frac{1}{M}\sum_{k=1}^{M}(x_{k,2}-\mu_2)^2 \cdots \frac{1}{M}\sum_{k=1}^{M}(x_{k,D}-\mu_D)^2\right] \in [1 \times D]$$

$$\hat{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \in [1 \times D]$$

$$y_i = \gamma \hat{x_i} + \beta \in [1 \times D]$$

Taking into consideration the method from paper, Notice that,

$$\frac{dl}{d\mu} = \cdots + \frac{dl}{d\sigma^2}\sum_{k=1}^{M} \frac{-2(x_i-\mu)}{M} = \cdots + \frac{dl}{d\sigma^2}\frac{-2}{M}\sum_{k=1}^{M}(x_i-\mu)$$

$$= \cdots + \frac{dl}{d\sigma^2}\frac{-2}{M}\left(\left(\sum_{k=1}^{M} x_i\right) - M\mu\right)$$

$$= \cdots + \frac{dl}{d\sigma^2}\frac{-2}{M} \times 0 = 0 \quad + \cdots$$

Then we can get

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\varepsilon}} + \left[\sum_{k=1}^{M} (x_k-\mu)\left(\frac{-1}{2}\right)(\sigma^2+\varepsilon)^{-3/2}\right] \cdot \frac{2(x_i-\mu)}{M}$$

$$+ \sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k} \frac{-1}{\sqrt{\sigma^2+\varepsilon}} \cdot \frac{1}{M}$$

$$= \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\varepsilon}} - \frac{(\sigma^2+\varepsilon)^{-\frac{3}{2}}}{M}\left[\sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}(x_k-\mu)\right](x_i-\mu)$$

$$- \frac{1}{M\sqrt{\sigma^2+\varepsilon}} \sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k} * \frac{(\sigma^2+\varepsilon)^{-3/2}}{M}\left[\sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}(x-\mu)\right](x_i-\mu)$$

$$= \frac{1}{M\sqrt{\sigma^2+\varepsilon}}\left[\sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}\frac{(x_k-\mu)}{(\sqrt{\sigma^2+\varepsilon})}\right]\frac{(x_i-\mu)}{\sqrt{\sigma^2+\varepsilon}}$$

$$= \frac{1}{M\sqrt{\sigma^2+\varepsilon}}\left[\sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}\hat{x}_k\right]\hat{x}_i$$

$$= \frac{1}{M\sqrt{\sigma^2+\varepsilon}} M\frac{\partial l}{\partial \hat{x}_i} - \frac{1}{M\sqrt{\sigma^2+\varepsilon}}\left[\sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}\hat{x}_k\right]\hat{x}_i$$

$$- \frac{1}{M\sqrt{\sigma^2+\varepsilon}} \sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}$$

$$= M\frac{\partial l}{\partial \hat{x}_i} - \left[\sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}\hat{x}_k\right]\hat{x}_i - \sum_{k=1}^{M} \frac{\partial l}{\partial \hat{x}_k}$$