Layer Normalization : Back propagation

$$x = \begin{matrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{matrix}$$

$$\mu = \begin{matrix} \dfrac{x_{11} + x_{12} + x_{13}}{3} \\[2ex] \dfrac{x_{21} + x_{22} + x_{23}}{3} \end{matrix}$$

$\downarrow$

$1 \times N$

$$\Rightarrow \mu = \frac{1}{D} \sum_{j=1}^{D} x_{1j} = \mu_1$$

$$\frac{1}{D} \sum_{j=1}^{D} x_{2j} = \mu_2$$

$$\sigma^2 = \frac{1}{3}\left((x_{11} - \mu_1)^2 + (x_{12} - \mu_1)^2 + (x_{13} - \mu_1)^2 \right)$$

$$\frac{1}{3}\left[(x_{21} - \mu_2)^2 + (x_{22} - \mu_2)^2 + (x_{23} - \mu_2)^2 \right]$$

$$\Rightarrow \sigma_1^2 = \frac{1}{D} \sum_{j=1}^{D} (x_{1j} - \mu_1)^2$$

$$\hat{x} = \begin{matrix} \dfrac{x_{11} - \mu_1}{(\sigma_1^2 + \epsilon)^{0.5}} & \dfrac{x_{12} - \mu_1}{(\sigma_1^2 + \epsilon)^{0.5}} & \dfrac{x_{13} - \mu_1}{(\sigma_1^2 + \epsilon)^{0.5}} \\[3ex] \dfrac{x_{21} - \mu_2}{(\sigma_2 + \epsilon)^{0.5}} & \dfrac{x_{22} - \mu_2}{(\sigma_2^2 + \epsilon)^{0.5}} & \dfrac{x_{23} - \mu_2}{(\sigma_2^2 + \epsilon)^{0.5}} \end{matrix}$$

$$y = \begin{matrix} \gamma_1 \hat{x}_{11} + \beta_1 & \gamma_2 \hat{x}_{12} + \beta_2 & \gamma_3 \hat{x}_{13} + \beta_3 \\ \gamma_1 \hat{x}_{21} + \beta_1 & \gamma_2 \hat{x}_{22} + \beta_2 & \gamma_3 \hat{x}_{23} + \beta_3 \end{matrix}$$

$$\frac{dL}{dy} = \begin{matrix} \dfrac{dL}{dy_{11}} & \dfrac{dL}{dy_{12}} & \dfrac{dL}{dy_{13}} \\[2ex] \dfrac{dL}{dy_{21}} & \dfrac{dL}{dy_{22}} & \dfrac{dL}{dy_{23}} \end{matrix}$$

$$\frac{dL}{d\gamma_1} = \sum_{i,j} \frac{dL}{dy_{ij}} \frac{dy_{ij}}{d\gamma_1}$$

$$= \cancel{\sum_{j=1}^{D} \frac{dL}{dy_{ij}} \frac{dy_{ij}}{}}$$

$$= \sum_{i=1}^{N} \frac{dL}{dy_{i1}} \cdot \frac{dy_{i1}}{d\gamma_1}$$

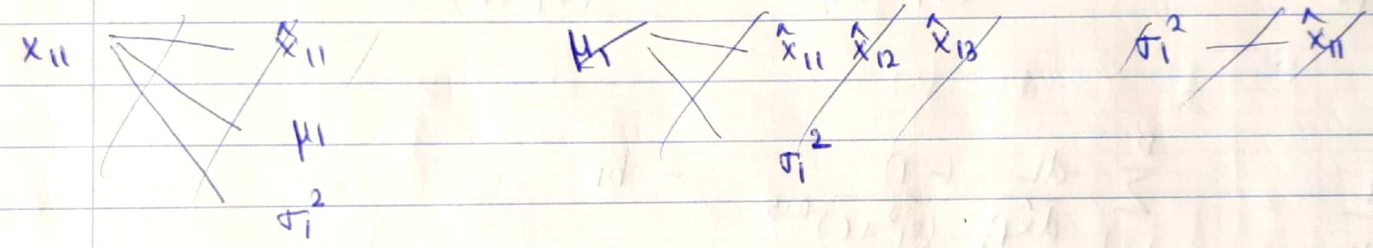
$$=$$

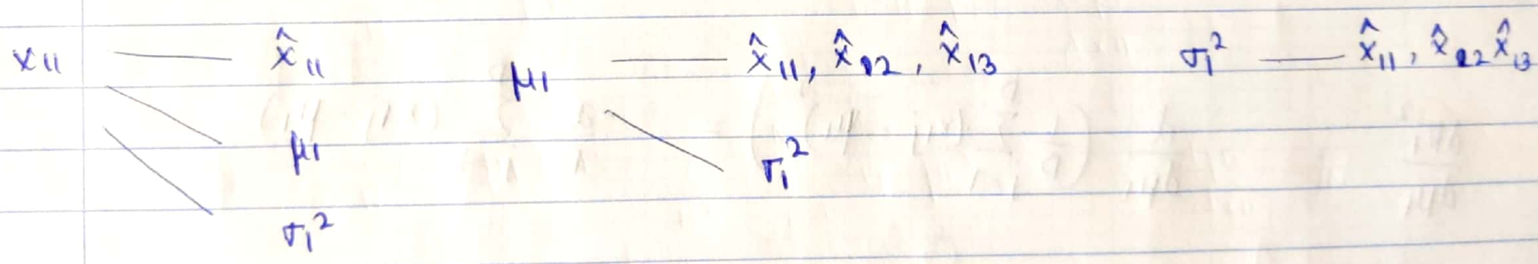
$$\frac{dL}{d\gamma_1} = \frac{dL}{dy_{11}} \cdot \hat{x}_{11} + \frac{dL}{dy_{21}} \hat{x}_{21} \Rightarrow \boxed{\frac{dL}{d\gamma} = \left(\frac{dL}{dy} \odot \hat{x}\right) \rightarrow \text{sum along axis=0}}$$

$$\frac{dL}{d\beta_1} = \sum_{i=1}^{N} \frac{dL}{dy_{i1}} \cdot \frac{dy_{i1}}{d\beta_1} = \frac{dL}{dy_{11}} + \frac{dL}{dy_{21}} \Rightarrow \boxed{\frac{dL}{d\beta} = \frac{dL}{dy} \rightarrow \text{sum along axis=0.}}$$

$$\frac{dL}{d\hat{x}_{11}} = \sum_{i,j} \frac{dL}{dy_{ij}} \frac{dy_{ij}}{d\hat{x}_{11}} = \frac{dL}{dy_{11}} \cdot \gamma_1$$

$$\boxed{\frac{dL}{d\hat{x}} = \left(\frac{dL}{dy} \odot \gamma\right)}$$





$$\frac{dL}{d\sigma_1^2} = \sum_{i,j} \frac{dL}{d\hat{x}_{ij}} \frac{d\hat{x}_{ij}}{d\sigma_1^2} = \frac{dL}{d\hat{x}_{11}} \frac{d\hat{x}_{11}}{d\sigma_1^2} + \frac{dL}{d\hat{x}_{12}} \frac{d\hat{x}_{12}}{d\sigma_1^2} + \frac{dL}{d\hat{x}_{13}} \frac{d\hat{x}_{13}}{d\sigma_1^2}$$

$$= \sum_{j=1}^{D} \frac{dL}{dx_{1j}} \frac{d\hat{x}_{1j}}{d\sigma_1^2}$$

$$\frac{d\hat{x}_{11}}{d\sigma_1^2} = (x_{11} - \mu_1)\left(\frac{-1}{2}\right)(\sigma_1^2 + \epsilon)^{-3/2}$$

$$\therefore \frac{dL}{d\sigma_1^2} = \frac{-1}{2}(\sigma_1^2 + \epsilon)^{-3/2} \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{1j}} (x_{1j} - \mu_1) =$$

$$\frac{dL}{d\sigma_1^2} = \frac{-1}{2(\sigma_1^2+\epsilon)} \sum_{j=1}^{D} \frac{(x_{1j} - \mu_1)}{(\sigma_1^2+\epsilon)^{0.5}} \frac{dL}{d\hat{x}_{1j}}$$

$$\boxed{\frac{dL}{d\sigma_1^2} = \frac{-1}{2(\sigma_1^2+\epsilon)} \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{1j}} \cdot \hat{x}_{1j}}$$

$$\frac{dL}{d\mu_1} = \sum_{i,j} \frac{dL}{d\hat{x}_{ij}} \frac{d\hat{x}_{ij}}{d\mu_1} = \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{1j}} \frac{d\hat{x}_{1j}}{d\mu_1}$$

$$\frac{d\hat{x}_{11}}{d\mu_1} = \frac{d}{d\mu_1} \frac{(x_{11}-\mu_1)}{(\sigma_1^2+\epsilon)^{0.5}} = \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}}$$

$$\frac{dL}{d\mu_1} = \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{1j}} \frac{(-1)}{(\sigma_1^2+\epsilon)^{0.5}} \leftarrow b_1$$

$$\frac{d\sigma_1^2}{d\mu_1} = \frac{d}{d\mu_1}\left(\frac{1}{D}\sum_{j=1}^{D}(x_{1j}-\mu_1)^2\right) = \frac{-2}{D}\sum_{j=1}^{D}(x_{1j}-\mu_1)$$

$$= 0$$

$$\frac{d\sigma_1^2}{d\mu_1} = 0$$

$$\therefore \frac{dL}{d\mu_1} = b_1 + \frac{dL}{d\sigma_1^2} \times \cancelto{0}{\frac{d\sigma_1^2}{d\mu_1}}$$

$$\therefore \boxed{\frac{dL}{d\mu_1} = \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}} \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{1j}}}$$

$$\frac{dL}{dx_{11}} = \sum_{ij} \frac{dL}{d\hat{x}_{ij}} \frac{d\hat{x}_{ij}}{dx_{11}} = \frac{dL}{d\hat{x}_{11}} \cdot \frac{d\hat{x}_{11}}{dx_{11}} = \frac{1}{(\sigma_i^2+\epsilon)^{0.5}} \frac{dL}{d\hat{x}_{11}} \quad - a_1$$

$$\cancel{\frac{dL}{}} \cancel{\frac{d\mu_i}{dx_{11}}} = \cancel{\sum_{ij} \frac{dL}{d\hat{x}_{ij}} \frac{d\hat{x}_{ij}}{d\mu_1}}$$

$$\frac{dL}{dx_{11}} = \sum_{j=1}^{N} \frac{dL}{d\mu_j} \frac{d\mu_j}{dx_{11}} = \frac{dL}{d\mu_1} \times \frac{d\mu_1}{dx_{11}}$$

$$= \left(\frac{-1}{(\sigma_i^2+\epsilon)^{0.5}} \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{ij}} \right) \cdot \frac{d\mu_1}{dx_{11}}$$

$$\frac{d\mu_1}{dx_{11}} = \frac{1}{D}$$

$$\therefore \frac{dL}{dx_{11}} + = \frac{-1}{D(\sigma_i^2+\epsilon)^{0.5}} \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{ij}} \quad — \quad a_2$$

$$\cancel{\frac{dL}{dx_{11}}} = \cancel{\sum_{j=1}^{D} \frac{dL}{d}} \qquad \frac{dL}{dx_{11}} = \sum_{i=1}^{N} \frac{dL}{d\sigma_i^2} \frac{d\sigma_i^2}{dx_{11}}$$

$$= \frac{dL}{d\sigma_i^2} \frac{d\sigma_i^2}{dx_{11}}$$

$$\frac{d\sigma_i^2}{dx_{11}} = \frac{d}{dx_{11}} \frac{1}{D} \sum_{j=1}^{D} (x_{ij}-\mu_1)^2 = \frac{2}{D}(x_{11}-\mu_1)$$

$$\therefore \frac{dL}{dx_{11}} = \frac{1}{(\sigma_i^2+\epsilon)^{0.5}} \frac{dL}{d\hat{x}_{11}} - \frac{1}{D(\sigma_i^2+\epsilon)^{0.5}} \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{ij}} - \frac{1}{2(\sigma_i^2+\epsilon)} \left(\sum_{j=1}^{D} \frac{dL}{d\hat{x}_{ij}} \hat{x}_{ij} \right) \cdot \frac{2}{D}(x_{11}-\mu_1)$$

$$\boxed{\therefore \frac{dL}{dx_{11}} = \frac{1}{(\sigma_i^2+\epsilon)^{0.5} D} \left[D \cdot \frac{dL}{d\hat{x}_{11}} - \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{ij}} - \hat{x}_{11} \sum_{j=1}^{D} \frac{dL}{d\hat{x}_{ij}} \hat{x}_{ij} \right]}$$