682 - Assignment - 2

Batch Normalization => alternative backward Derivation

required :- $\frac{\partial l}{\partial x_i}$, $\frac{\partial l}{\partial \gamma}$, $\frac{\partial l}{\partial \beta}$ with $\frac{\partial l}{\partial \hat{x}_i}$, $\frac{\partial l}{\partial \mu_B}$, $\frac{\partial l}{\partial \sigma_B^2}$

μ_B - Mean of $(1 \times D)$ sized batch

σ_B^2 - variance $\hat{x}_1$ - normalized i/p

$$\hat{x}_i = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \qquad y = \gamma \hat{x}_i + \beta$$

$$\Rightarrow \frac{\partial y}{\partial \hat{x}_i} = \gamma \qquad \frac{\partial \hat{x}_i}{\partial \mu_B} \Rightarrow \frac{-1}{\sqrt{\sigma^2 + \epsilon}} \text{ (chain rule)}$$

accordingly σ^2 = variance = $\frac{1}{m} \sum_{i=1}^{m} (\hat{x}_i - \mu_B)^2$

$$\frac{\partial \sigma_B^2}{\partial \mu_B} = \frac{1}{m} \sum_{i=1}^{m} -2 (\hat{x}_i - \mu_B)$$

By chain rule $\Rightarrow \frac{\partial l}{\partial \sigma_B^2} = \frac{\partial l}{\partial \hat{x}_i} \times \frac{\partial \hat{x}_i}{\partial \sigma_B^2}$

We know that $\hat{x}_i = \frac{(\hat{x}_i - \mu_B)}{\sqrt{\sigma_B^2 + \epsilon}}$

$$\Rightarrow \frac{\partial \hat{x}_i}{\partial \sigma_B^2} \Rightarrow \frac{\partial \hat{x}_i}{\partial \sigma_B^2} (\hat{x}_i - \mu_B)(\sigma_B^2 + \epsilon)^{-1/2}$$

$$\Rightarrow -\frac{1}{2} \sum_{i=1}^{m} (\hat{x}_i - \mu_B)(\sigma_B^2 + \epsilon)^{-3/2} \qquad \text{— (1)}$$

We require $\frac{\partial l}{\partial \mu_B} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu_B} + \frac{\partial l}{\partial \sigma_B^2} \frac{\partial \sigma_B^2}{\partial \mu_B}$

$$\Rightarrow \left(\sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \left(\frac{\partial l}{\partial \sigma_B^2} \cdot \frac{1}{m} \sum_{i=1}^{m} -2(\hat{x}_i - \mu_B) \right)$$

$$\Rightarrow \left(\sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_B^2+\epsilon}}\right) + \left(-2\frac{\partial l}{\partial \sigma_B^2}\left[\frac{1}{m}\sum_{i=1}^{m} \hat{x}_i - \frac{1}{m}\sum_{i=1}^{m} \mu_B\right)\right)$$

$$\Rightarrow \left(\sum_{i=1}^{m} -\frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2+\epsilon}}\right) + \left(-2\frac{\partial l}{\partial \sigma_B^2} \cdot \left(\mu_B - \frac{\mu_B \cdot m}{m}\right)\right)$$

$$\Rightarrow \left(\sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \frac{-1}{\sqrt{\sigma_B^2+\epsilon}}\right) \text{ --- } (2)$$

Now $$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial l}{\partial \mu_B} \cdot \frac{\partial \mu_B}{\partial x_i} + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{\partial \sigma_B^2}{\partial x_i}$$

required

$$\frac{\partial \hat{x}_i}{\partial x_i} = \left(\sqrt{\sigma^2+\epsilon}\right)^{-1} \text{ \& } \frac{\partial \mu_B}{\partial x_i} = \frac{1}{m}$$

$$\frac{\partial \sigma_B^2}{\partial x_i} = \frac{2}{m}(x_i - \mu_B)$$

required $$\Rightarrow \frac{\partial l}{\partial x_i} = \left(\frac{\partial l}{\partial \hat{x}_i} \cdot (\sigma_B^2+\epsilon)^{-1/2}\right) + \left(\frac{\partial l}{\partial \mu_B} \cdot \frac{1}{m}\right) + \left(\frac{\partial l}{\partial \sigma_B^2} \cdot \frac{2}{m}(x_i - \mu_B)\right)$$

From (1), (2)

$$\Rightarrow \frac{\partial l}{\partial x_i} = \left(\frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{(\sigma_B^2+\epsilon)^{1/2}}\right) + \left(\frac{1}{m}\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \frac{-1}{\sqrt{\sigma_B^2+\epsilon}}\right)$$

$$- \left(\frac{1}{2}\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}(x_j - \mu_B)(\sigma_B^2+\epsilon)^{-1/2} \cdot \frac{2}{m}(x_i - \mu_B)\right)$$

$$\Rightarrow \left(\frac{\partial l}{\partial \hat{x}_i} \cdot (\sigma_B^2+\epsilon)^{-1/2}\right) - \left((\sigma_B^2+\epsilon)^{-1/2} \cdot \frac{1}{m}\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right) +$$

$$+ \left((\sigma_B^2+\varepsilon)^{-1/2} \cdot \frac{1}{m} \cdot \hat{x}_i \sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \hat{x}_j \right)$$

$$\Rightarrow \textcircled{1} \quad \boxed{\frac{\partial l}{\partial x_i} = \frac{(\sigma_B^2+\varepsilon)^{-1/2}}{m} \left[m \frac{\partial l}{\partial \hat{x}_i} - \sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} - \hat{x}_i \sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \hat{x}_j \right]}$$

Hence Derived $\frac{\partial l}{\partial x_i}$

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial y_i} \cdot \frac{\partial y_i}{\partial \beta} \Rightarrow \sum_{i=1}^{m} \frac{\partial l}{\partial y_i}$$

$$\Rightarrow \frac{\partial l}{\partial \hat{x}_i} = \frac{\partial l}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} \Rightarrow \frac{\partial l}{\partial y_i} \cdot \gamma$$

$$\frac{\partial l}{\partial \gamma} = \frac{\partial l}{\partial y_i} \cdot \frac{\partial y_i}{\partial \gamma} \Rightarrow \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \cdot \hat{x}_i$$

$$\textcircled{2} \quad \boxed{\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \hat{x}_i} \qquad \textcircled{3} \quad \boxed{\frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i}}$$

$$\textcircled{4} \quad \boxed{\frac{\partial l}{\partial y_i} = dout}$$