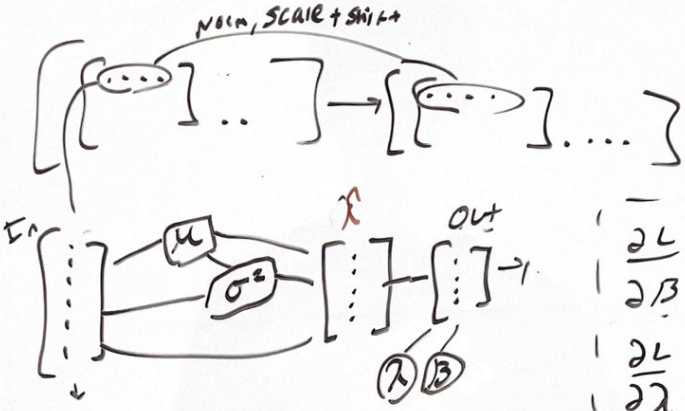
Deriving Layer norm.



$$\frac{\partial L}{\partial \beta} = \left(\frac{\partial L}{\partial out}\right)$$

$$\frac{\partial L}{\partial \lambda} = \left(\frac{\partial L}{\partial out} \cdot raw\right) \cdot sum$$

$$\frac{\partial L}{\partial \hat{x}_i} = \frac{\partial L}{\partial out_i} \cdot \lambda$$

$$\mu = \frac{\sum_n x_n}{n}$$

$$\sigma^2 = \left(\frac{\sum_n (x_n - \mu)^2}{n-1}\right)$$

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2+\epsilon}}$$

$$y_i = \hat{x}_i \cdot \lambda + \beta$$

$$\boxed{\frac{\partial L}{\partial raw_i} = \lambda \cdot \frac{\partial L}{\partial out_i}}$$

$$\frac{\partial L}{\partial \sigma^2} = \sum_i \frac{\partial L}{\partial raw_i} \cdot \frac{\partial raw_i}{\partial \sigma^2} \rightarrow \sum_i \lambda \cdot \frac{\partial L}{\partial out_i} \cdot \frac{d}{d\sigma^2}\left(\frac{x_i - \mu}{\sqrt{\sigma^2+\epsilon}}\right) \rightarrow \sum_i \lambda \cdot \frac{\partial L}{\partial out_i} \cdot -\frac{1}{2}\left(\frac{x_i - \mu}{(\sigma^2+\epsilon)^{3/2}}\right)$$

$$\frac{\hat{x}_i}{(\sigma^2+\epsilon)^1}$$

$$= \boxed{\sum_i \beta \cdot \frac{\partial L}{\partial out_i} \cdot -\frac{1}{2} \frac{raw_i}{\sqrt{\sigma^2+\epsilon}}}$$

$$\frac{\partial L}{\partial \mu} = \underbrace{\frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu^2}} + \underbrace{\sum_i \frac{\partial L}{\partial raw_i} \cdot \frac{\partial raw_i}{\partial L}} \rightarrow \boxed{\sum_i \frac{-1}{\sqrt{\sigma^2+\epsilon}} \cdot \frac{\partial L}{\partial out_i} \cdot \lambda = \frac{\partial L}{\partial \mu}}$$

$$\cancel{\frac{\partial L}{\partial \sigma^2} \cdot -2 \sum_i \frac{x_i - \mu}{n-1}} \xrightarrow{} 0 = 0$$

Note:

$$\sum \frac{x_i - \mu}{\sigma + \epsilon} \rightarrow \frac{1}{\sqrt{\sigma+\epsilon}} \sum \frac{x_i - \mu}{\sqrt{\sigma+\epsilon}}$$

and,

$$\sum x_i - \mu = 0$$

and,

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2+\epsilon}}$$

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial x_i} + \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i}$$

$$\frac{1}{n} \cdot \frac{\partial L}{\partial \mu} + 2 \cdot \frac{x_i - \mu}{n-1} \cdot \frac{\partial L}{\partial \sigma^2} + \frac{1}{\sqrt{\sigma^2+\epsilon}} \cdot \frac{\partial L}{\partial \hat{x}_i} \rightarrow$$

$$\frac{1}{n} \sum_i \frac{-1}{\sqrt{\sigma^2+\epsilon}} \cdot \frac{\partial L}{\partial out_i} \cdot \lambda$$

$$+ \frac{x_i - \mu}{(n-1)\sqrt{\sigma^2+\epsilon}} \sum_i \frac{\hat{x}_i}{\sqrt{\sigma^2+\epsilon}} \cdot \frac{\partial L}{\partial out_i} \cdot -\frac{1}{2} \cdot \lambda$$

$$+ \frac{1}{\sqrt{\sigma^2+\epsilon}} \cdot \frac{\partial L}{\partial out_i} \cdot \lambda$$

$$= \boxed{\frac{\lambda \cdot n^{-1}}{\sqrt{\sigma^2+\epsilon}} \left(\frac{\partial L}{\partial out_i} \cdot n - \sum_i \frac{\partial L}{\partial out_i} - \frac{\hat{x}_i \cdot n}{n-1} \sum_i \hat{x}_i \cdot \frac{\partial L}{\partial out_i}\right)} = \frac{\partial L}{\partial x}$$