基本公式

$$y = \frac{x-\mu}{\sqrt{\sigma+\epsilon}} \odot \gamma + \beta$$

其中:

- 均值

$$\mu = \frac{1}{D}\sum_{k=1}^{D} x_i$$

- 方差

$$\sigma^2 = \frac{1}{D}\sum_{k=1}^{D} (x_i-\mu)^2$$

- 计算中间值,用于反向传播

$$rstd = \frac{1}{\sqrt{\sigma^2+\epsilon}}$$

前向传播公式

对于每个输入张量 x,层归一化对每个样本 $x_i = [x_{i1}, x_{i2}, ..., x_{iD}]$ 进行操作。

1. 计算均值 (Mean)

对每个样本 i,计算其所在特征的均值 μ_i:

$$\mu_i = \frac{1}{D}\sum_{j=1}^{D} x_{ij}$$

2. 计算方差 (Variance)

对于每个样本 i, 计算其所有特征的方差 σ_i^2:

$$\sigma_i^2 = \frac{1}{D}\sum_{j=1}^{D} (x_{ij}-\mu_i)^2$$

3. 归一化(Normalization)

使用均值和方差对每个特征进行归一化，得到 $\hat{x}_{ij}$：

$$\hat{x}_{ij} = \frac{x_{ij}-\mu_i}{\sqrt{\sigma_i^2+\epsilon}}$$

其中 ϵ 是一个很小的常数 (例如 $\frac{1}{10^5}$),用于防止除以零。

4. 缩放和平移 (Scale and Shift)

引入可学习参数 $\gamma \in \mathbb{R}^D$ (缩放因子) 和 $\beta \in \mathbb{R}^D$ (平移因子), 对归一化后的特征进行缩放和平移:

$$y_{ij} = \gamma_j \hat{x}_{ij} + \beta_j$$

这里，γ_j 和 β_j 是对应于第 j 个特征的参数。

反向传播公式推导

反向传播的目标是计算损失函数 L 对输入 x 以及可学习参数 γ 和 β 的梯度。

假设我们已经得到了损失函数 L 对输出 y 的梯度 $\frac{\partial L}{\partial y_{ij}}$.

1. 对 γ 和 β 的梯度:

根据 $y_{ij} = \gamma_j \hat{x}_{ij} + \beta_j$,我们有:

$$\frac{\partial L}{\partial \gamma_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial \gamma_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \hat{x}_{ij}$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \cdot 1$$

2. 对 $\hat{x}_{ij}$ 的梯度:

$$\frac{\partial L}{\partial \hat{x}_{ij}} = \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial \hat{x}_{ij}} = \frac{\partial L}{\partial y_{ij}} \gamma_j$$

3. 对 μ_i 和 σ_i^2 的梯度:

这个部分比较复杂，因为 μ_i 和 σ_i^2 都是 x_{ij} 的函数，并且 $\hat{x}_{ij}$ 也依赖他们。

我们有:

$$\hat{x}_{ij} = \frac{x_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

因此:

$$\frac{\partial \hat{x}_{ij}}{\partial \mu_i} = -\frac{1}{\sqrt{\sigma_i^2 + \epsilon}}$$

$$\frac{\partial \hat{x}_{ij}}{\partial \sigma_i^2} = -\frac{1}{2} \left(x_{ij} - \mu_i\right) \left(\sigma_i^2 + \epsilon\right)^{-\frac{3}{2}}$$

因此,根据链式法则，$\frac{\partial L}{\partial \mu_i}$ 和 $\frac{\partial L}{\partial \sigma_i^2}$ 的计算需要对所有 j 求和:

$$\frac{\partial L}{\partial \mu_i} = \sum_{j=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}} \frac{\partial \hat{x}_{ij}}{\partial \mu_i}$$

$$\frac{\partial L}{\partial \sigma_i^2} = \sum_{j=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}} \frac{\partial \hat{x}_{ij}}{\partial \sigma_i^2}$$

代入 $\frac{\partial \hat{x}_{ij}}{\partial \mu_i}$ 和 $\frac{\partial \hat{x}_{ij}}{\partial \sigma_i^2}$ 的表达式:

$$\frac{\partial L}{\partial \mu_i} = \sum_{j=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}} \left(-\frac{1}{\sqrt{\sigma_i^2+\epsilon}}\right) = \left(-\frac{1}{\sqrt{\sigma_i^2+\epsilon}}\right) \sum_{j=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}}$$

$$\frac{\partial L}{\partial \sigma_i^2} = \sum_{j=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}} \left(-\frac{1}{2}\left(x_{ij}-\mu_i\right)\left(\sigma_i^2+\epsilon\right)^{-\frac{3}{2}}\right) = -\frac{1}{2\left(\sigma_i^2+\epsilon\right)^{\frac{3}{2}}} \sum_{j=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}} \left(x_{ij}-\mu_i\right)$$

4. 对 x_{ij} 的梯度

这里比较复杂，因为 x_{ij} 不仅直接影响 $\hat{x}_{ij}$，还通过 μ 和 σ_i^2 间接影响 $\hat{x}_{ij}$

$$\frac{\partial L}{\partial x_{ij}} = \frac{\partial L}{\partial \hat{x}_{ij}} \frac{\partial \hat{x}_{ij}}{\partial x_{ij}} + \frac{\partial L}{\partial \mu_i} \frac{\partial \mu_i}{\partial x_{ij}} + \frac{\partial L}{\partial \sigma_i^2} \frac{\partial \sigma_i^2}{\partial x_{ij}}$$

我们分别计算这些偏导数:

- $\frac{\partial \hat{x}_{ij}}{\partial x_{ij}} = \frac{1}{\sqrt{\sigma_i^2+\epsilon}}$
- $\frac{\partial \mu_i}{\partial x_{ij}} = \frac{\partial\left(\frac{1}{D}\sum_{j=1}^{D} x_{ij}\right)}{\partial x_{ij}} = \frac{1}{D}$
- $\frac{\partial \sigma_i^2}{\partial x_{ij}}$，注意：$\sigma_i^2$ 是 x_{ij}的函数，也是 μ_i 的函数，而 μ_i 是 x_{ij} 的函数，因此:

$$\frac{\partial \sigma_i^2}{\partial x_{ij}} = \frac{\partial\left(\frac{1}{D}\sum_{k=1}^{D}\left(x_{ik}-\mu_i\right)^2\right)}{\partial x_{ij}}$$

$$= \frac{1}{D}\left(\frac{\partial\left(\sum_{k=1}^{D}\left(x_{ik}^2-2x_{ik}\mu_i+\mu_i^2\right)\right)}{\partial x_{ij}}\right)$$

$$= \frac{1}{D}\left(\frac{\partial \sum_{k=1}^{D} x_{ik}^2}{\partial x_{ij}} - \frac{2\partial \sum_{k=1}^{D} x_{ik}\mu_i}{\partial x_{ij}} + \frac{\partial \sum_{k=1}^{D} \mu_i^2}{\partial x_{ij}}\right)$$

$$= \frac{1}{D}\left[2x_{ij} - 2\left(\frac{\partial \sum_{k=1}^{D} x_{ik}}{\partial x_{ij}}\mu_i + \frac{\partial \mu_i}{\partial x_{ij}} \sum_{k=1}^{D} x_{ik}\right) + \left(2\sum_{k=1}^{D} \mu_i \frac{\partial \mu_i}{\partial x_{ij}}\right)\right]$$

$$= \frac{1}{D}\left[2x_{ij} - 2\left(\mu_i + \frac{1}{D} D\mu_i\right) + 2\sum_{k=1}^{D} \mu_i \frac{1}{D}\right]$$

$$= \frac{1}{D}\left[2x_{ij} - 2\left(\mu_i+\mu_i\right) + 2\mu_i\right]$$

$$= \frac{2}{D}\left(x_{ij}-\mu_i\right)$$

或者，更加简洁的推导:

$$\frac{\partial \sigma_i^2}{\partial x_{ij}} = \frac{\partial\left(\frac{1}{D}\sum_{k=1}^{D}\left(x_{ik}-\mu_i\right)^2\right)}{\partial x_{ij}}$$

$$
\begin{aligned}
&= \frac{1}{D} \sum_{k=1}^{D} \frac{\partial \left(x_{ik} - \mu_i\right)^2}{\partial x_{ij}} \\
&= \frac{1}{D} \sum_{k=1}^{D} 2 \left(x_{ik} - \mu_i\right) \left(\frac{\partial x_{ik}}{\partial x_{ij}} - \frac{\partial \mu_i}{\partial x_{ij}}\right) \\
&= \frac{1}{D} \sum_{k=1}^{D} 2 \left(x_{ik} - \mu_i\right) \left(\delta_{jk} - \frac{\partial \mu_i}{\partial x_{ij}}\right) \\
&= \frac{2}{D} \left[\sum_{k=1}^{D} \left(x_{ik} - \mu_i\right) \delta_{jk} - \sum_{k=1}^{D} \left(x_{ik} - \mu_i\right) \frac{\partial \mu_i}{\partial x_{ij}}\right] \\
&= \frac{2}{D} \left[\left(x_{ij} - \mu_i\right) - \frac{1}{D} \sum_{k=1}^{D} \left(x_{ik} - \mu_i\right)\right]
\end{aligned}
$$

计算 $\sum_{k=1}^{D} \left(x_{ik} - \mu_i\right)$：

$$
\sum_{k=1}^{D} \left(x_{ik} - \mu_i\right) = \sum_{k=1}^{D} x_{ik} - \sum_{k=1}^{D} \mu_i = D\mu_i - D\mu_i = 0
$$

因此:

$$
\frac{\partial \sigma_i^2}{\partial x_{ij}} = \frac{2}{D} \left[\left(x_{ij} - \mu_i\right) - \frac{1}{D} \sum_{k=1}^{D} \left(x_{ik} - \mu_i\right)\right] = \frac{2}{D} \left(x_{ij} - \mu_i\right)
$$

现在，将所有项代入 $\frac{\partial L}{\partial x_{ij}}$ 的公式:

$$
\begin{aligned}
\frac{\partial L}{\partial x_{ij}} &= \frac{\partial L}{\partial \hat{x}_{ij}} \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} + \left(-\frac{1}{\sqrt{\sigma_i^2 + \epsilon}} \sum_{j=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}}\right) \frac{1}{D} + \left(-\frac{1}{2\left(\sigma_i^2 + \epsilon\right)^{\frac{3}{2}}} \sum_{k=1}^{D} \frac{\partial L}{\partial \hat{x}_{ik}} \left(x_{ik} - \mu_i\right)\right) \frac{2}{D} \left(x_{ij}\right. \\
&= \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} \left[\frac{\partial L}{\partial \hat{x}_{ij}} - \frac{1}{D} \sum_{k=1}^{D} \frac{\partial L}{\partial \hat{x}_{ik}} - \frac{x_{ij} - \mu_i}{D\left(\sigma_i^2 + \epsilon\right)} \sum_{k=1}^{D} \frac{\partial L}{\partial \hat{x}_{ik}} \left(x_{ik} - \mu_i\right)\right]
\end{aligned}
$$

其中:

$$
\hat{x}_{ij} = \frac{x_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}
$$

$$
\hat{x}_{ik} = \frac{x_{ik} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}
$$

所以:

$$
\frac{\partial L}{\partial x_{ij}} = \frac{1}{D} \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} \left(D \frac{\partial L}{\partial \hat{x}_{ij}} - \sum_{k=1}^{D} \frac{\partial L}{\partial \hat{x}_{ik}} - \hat{x}_{ij} \sum_{k=1}^{D} \hat{x}_{ik} \frac{\partial L}{\partial \hat{x}_{ik}}\right)
$$

最终，所以的梯度公式如下:

$$\frac{\partial L}{\partial \gamma_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial \gamma_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \hat{x}_{ij}$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_{ij}} \cdot 1$$

$$\frac{\partial L}{\partial x_{ij}} = \frac{1}{D} \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} \left(D \frac{\partial L}{\partial \hat{x}_{ij}} - \sum_{k=1}^{D} \frac{\partial L}{\partial \hat{x}_{ik}} - \hat{x}_{ij} \sum_{k=1}^{D} \hat{x}_{ik} \frac{\partial L}{\partial \hat{x}_{ik}} \right)$$

为了提高计算效率，我们在前向传播时，需要缓存中间计算结果:

$$mean_i = \frac{1}{D} \sum_{k=1}^{D} x_{ik}$$

$$rstd_i = \frac{1}{\sqrt{\sigma_i^2 + \epsilon}}$$

从另外的视角看反向传播公式

- 根据链式法则:

$$\frac{\partial L}{\partial x_{ij}} = \frac{\partial L}{\partial \hat{x}_{ij}} \frac{\partial \hat{x}_{ij}}{\partial x_{ij}} + \frac{\partial L}{\partial \mu_i} \frac{\partial \mu_i}{\partial x_{ij}} + \frac{\partial L}{\partial \sigma_i^2} \frac{\partial \sigma_i^2}{\partial x_{ij}} + \frac{\partial L}{\partial \sigma_i^2} \frac{\partial \sigma_i^2}{\partial \mu_i} \frac{\partial \mu_i}{\partial x_{ij}}$$

因为:

$$\frac{\partial \sigma_i^2}{\partial \mu_i} = \frac{2}{D} \sum_{k=1}^{D} (\mu_i - x_{ik}) = 2 \left(\frac{1}{D} D \mu_i - \frac{1}{D} \sum_{k=1}^{D} x_{ik} \right) = 2 (\mu_i - \mu_i) = 0$$

有一般性的结论：由于方差的数学性质，方差对均值的导数恰好为零。
可以从最小二乘的角度看，方差实际上是:

$$\sigma^2 = \min_c \frac{1}{D} \sum_{i=1}^{D} (x_i - c)^2$$

在最优点处，目标函数对参数的导数必须为零:

$$\frac{\partial \sigma^2}{\partial c} = 0$$

解得:

$$c = \frac{1}{D} \sum_{i=1}^{D} x_i$$

这正是均值，因而，方差的最优化解是在导数为零的时候取得。