

Derivative of BatchNorm and LayerNorm operation w.r.t X

Kadulin V.

October 16, 2022

Let $Y = \gamma\hat{X} + \beta$, where $\hat{X} \in \mathbb{R}^{N \times D}$, $\gamma \in \mathbb{R}^D$, $\beta \in \mathbb{R}^D$, $\hat{X} = \frac{X - \mu}{\sigma}$. Let's start from BatchNorm case, where $\mu = \frac{1}{N} \sum_{i=1}^N x_i$, $v = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$, $\sigma = \sqrt{v + \epsilon}$, $\epsilon \in \mathbb{R}$ is a small scalar to omit zero division.

Let $f : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}$ is a differentiable scalar function. We also know $\frac{\partial f}{\partial Y}$. Derive $\frac{\partial f}{\partial X}$.

Notice that

- y_i depends on x_j , where $1 \leq j \leq N$
- there is only per-column dependency: y_{ij} depends on x_{kj} , where $1 \leq k \leq N$

Consider first column of X as x .

$$\begin{aligned} y &= \gamma\hat{x} + \beta \\ y, x, \hat{x} &\in \mathbb{R}^N \\ \gamma, \beta, \mu, v, \sigma &\in \mathbb{R} \end{aligned}$$

Let's consider j -th element of x as x_j .

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^N \frac{\partial f}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial x_j}$$

Let's derive expressions for building blocks:

$$\begin{aligned}\frac{\partial y}{\partial \hat{x}_i} &= \gamma \\ \frac{\partial \hat{x}_i}{\partial x_j} &= \frac{\partial(x_i - \mu)}{\partial x_j} (v + \epsilon)^{-\frac{1}{2}} + (x_i - \mu) \left(-\frac{1}{2}\right) (v + \epsilon)^{-\frac{3}{2}} \frac{\partial(v + \epsilon)}{\partial x_j} \\ \frac{\partial \mu}{\partial x_j} &= \frac{1}{N} \\ \frac{\partial x_i}{\partial x_j} &= \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}\end{aligned}$$

Let's derive $\frac{\partial v}{\partial x_j}$.

1 Forward-mode differentiation

Consider i -th component of sum in v as y_i .

$$\frac{\partial y_i}{\partial x_j} = \begin{cases} 2(x_i - \mu)(1 - \frac{1}{N}) = -\frac{2}{N}(x_i - \mu) + 2(x_i - \mu) & i = j \\ -\frac{2}{N}(x_i - \mu) & i \neq j \end{cases}$$

So, $\frac{\partial v}{\partial x_j} = \frac{1}{N}(-\frac{2}{N} \sum_{i=1}^N (x_i - \mu) + 2(x_j - \mu))$.

Note, that $\sum_{i=1}^N (x_i - \mu) = \sum_{i=1}^N x_i - N\mu = N\mu - N\mu = 0$.

So, $\frac{\partial v}{\partial x_j} = \frac{2}{N}(x_j - \mu)$.

Now we can expand $\frac{\partial \hat{x}_i}{\partial x_j}$. Let's do it in case $i \neq j$:

$$\begin{aligned}\frac{\partial \hat{x}_i}{\partial x_j} &= -\frac{1}{N}(v + \epsilon)^{-\frac{1}{2}} + (x_i - \mu) \left(-\frac{1}{2}\right) (v + \epsilon)^{-\frac{3}{2}} \frac{2}{N}(x_j - \mu) = \\ &\quad -\frac{1}{N}(v + \epsilon)^{-\frac{1}{2}} (1 + (x_i - \mu)(x_j - \mu)(v + \epsilon)^{-1})\end{aligned}$$

Note that the only difference of case $i = j$ is the first multiplier of the first term: it's $1 - \frac{1}{N}$ instead of $-\frac{1}{N}$.

Now we have all to derive $\frac{\partial f}{\partial x_j}$:

$$\begin{aligned}
\frac{\partial f}{\partial x_j} &= \sum_{i=1}^N \frac{\partial f}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial x_j} = \gamma \sum_{i=1}^N \frac{\partial f}{\partial y_i} \frac{\partial \hat{x}_i}{\partial x_j} = \\
\gamma \left(\sum_{i=1}^N -\frac{1}{N} (v + \epsilon)^{-\frac{1}{2}} (1 + (x_i - \mu)(x_j - \mu)(v + \epsilon)^{-1}) \frac{\partial f}{\partial y_i} + (v + \epsilon)^{-\frac{1}{2}} \frac{\partial f}{\partial y_j} \right) &= \\
\gamma (v + \epsilon)^{-\frac{1}{2}} \left(-\frac{1}{N} \sum_{i=1}^N (1 + (x_i - \mu)(x_j - \mu)(v + \epsilon)^{-1}) \frac{\partial f}{\partial y_i} + \frac{\partial f}{\partial y_j} \right) &= \\
\gamma (v + \epsilon)^{-\frac{1}{2}} \left(-\frac{1}{N} \sum_{i=1}^N \frac{\partial f}{\partial y_i} - (x_j - \mu)(v + \epsilon)^{-1} \frac{1}{N} \sum_{i=1}^N (x_i - \mu) \frac{\partial f}{\partial y_i} + \frac{\partial f}{\partial y_j} \right)
\end{aligned}$$

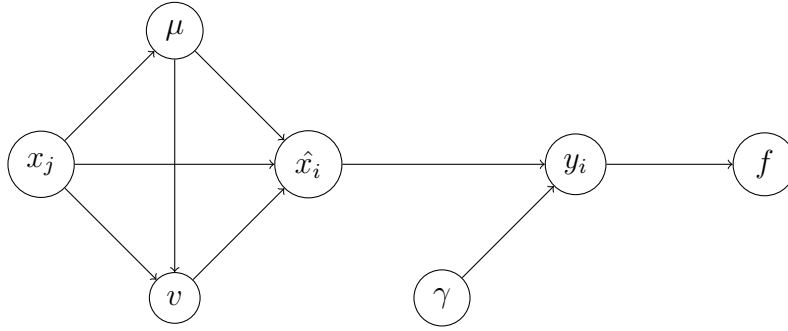
Knowing that, it's easy to write expression in matrix form:

$$\frac{\partial f}{\partial X} = \gamma (v + \epsilon)^{-\frac{1}{2}} \left(-\frac{1}{N} \sum_{i=1}^N \frac{\partial f}{\partial y_i} - (X - \mu)(v + \epsilon)^{-1} \frac{1}{N} \sum_{i=1}^N (x_i - \mu) \frac{\partial f}{\partial y_i} + \frac{\partial f}{\partial Y} \right)$$

Here $x, y, v, \mu, \gamma \in \mathbb{R}^D$.

2 Backward-mode differentiation

Let's draw a computation graph:



$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^N \frac{\partial f}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \left(\frac{\partial \hat{x}_i}{\partial x_j} + \frac{\partial \hat{x}_i}{\partial v} \frac{\partial v}{\partial x_j} + \left(\frac{\partial \hat{x}_i}{\partial \mu} + \frac{\partial \hat{x}_i}{\partial v} \frac{\partial v}{\partial \mu} \right) \frac{\partial \mu}{\partial x_j} \right)$$

$$\begin{aligned}
\frac{\partial y_i}{\partial \hat{x}_i} &= \gamma \\
\frac{\partial \hat{x}_i}{\partial x_j} &= \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \\
\frac{\partial \hat{x}_i}{\partial v} &= (x_i - \mu)(-\frac{1}{2})(v + \epsilon)^{-\frac{3}{2}} \\
\frac{\partial v}{\partial x_j} &= \frac{2}{N}(x_j - \mu) \\
\frac{\partial \hat{x}_i}{\partial \mu} &= -(v + \epsilon)^{-\frac{1}{2}} \\
\frac{\partial v}{\partial \mu} &= -\frac{2}{N} \sum_{i=1}^N (x_i - \mu) = 0 \\
\frac{\partial \mu}{\partial x_j} &= \frac{1}{N}
\end{aligned}$$

Now let's derive $\frac{\partial f}{\partial x_j}$:

$$\begin{aligned}
\frac{\partial f}{\partial x_j} &= \gamma \left(\sum_{i=1}^N \frac{\partial f}{\partial y_i} \left(-(x_i - \mu)(v + \epsilon)^{-\frac{3}{2}} \frac{1}{N}(x_j - \mu) - \frac{1}{N}(v + \epsilon)^{-\frac{1}{2}} \right) + (v + \epsilon)^{-\frac{1}{2}} \frac{\partial f}{\partial y_j} \right) = \\
&\gamma(v + \epsilon)^{-\frac{1}{2}} \left(-\frac{1}{N} \sum_{i=1}^N \frac{\partial f}{\partial y_i} - (x_j - \mu)(v + \epsilon)^{-1} \frac{1}{N} \sum_{i=1}^N (x_i - \mu) \frac{\partial f}{\partial y_i} + \frac{\partial f}{\partial y_j} \right)
\end{aligned}$$

Note that result is the same as in previous approach.

3 LayerNorm

This operation differs from BatchNorm in a single aspect: μ and v are calculated by columns instead of rows. This implies the following changes:

$$\begin{aligned}
\frac{\partial y_i}{\partial \hat{x}_i} &= \gamma_i \\
\frac{\partial \mu}{\partial x_j} &= \frac{1}{D}
\end{aligned}$$

So, the equation for LayerNorm operation looks as follows:

$$\frac{\partial f}{\partial X} = (v + \epsilon)^{-\frac{1}{2}} \left(-\frac{1}{D} \sum_{i=1}^D \gamma_i \frac{\partial f}{\partial y_i} - (X - \mu)(v + \epsilon)^{-1} \frac{1}{D} \sum_{i=1}^D \gamma_i (x_i - \mu) \frac{\partial f}{\partial y_i} + \gamma_j \frac{\partial f}{\partial Y} \right)$$