Derivative of BatchNorm operation w.r.t X

Kadulin V.

October 13, 2022

Let $Y = \gamma\hat{X} + \beta$, where $\hat{X} \in \mathbb{R}^{N\times D}$, $\gamma \in \mathbb{R}^D$, $\beta \in \mathbb{R}^D$, $\hat{X} = \frac{X-\mu}{\sigma}$, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\sigma = \sqrt{v+\epsilon}$, $v = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$, $\epsilon \in \mathbb{R}$ is a small scalar to omit zero division. Let $f : \mathbb{R}^{N\times D} \to \mathbb{R}$ 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 $1 \le j \le N$
- there is only per-column dependency: y_{ij} depends on x_{kj}, where $1 \le k \le N$

Consider first column of X as x.

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

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:

$\frac{\partial y}{\partial \hat{x}_i} = \gamma$

$\frac{\partial \hat{x}_i}{\partial x_j} = \frac{\partial(x_i-\mu)}{\partial x_j}(v+\epsilon)^{-\frac{1}{2}} + (x_i-\mu)(-\frac{1}{2})(v+\epsilon)^{-\frac{3}{2}}\frac{\partial(v+\epsilon)}{\partial x_j}$

$\frac{\partial \mu}{\partial x_j} = \frac{1}{N}$

$\frac{\partial x_i}{\partial x_j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$
Let's derive $\frac{\partial v}{\partial x_j}$.
Consider i-th component of sum in v as y_i.
$\frac{\partial y_i}{\partial x_j} = \begin{cases} 2(x_i - \mu)(1 - \frac{1}{N}) = -\frac{2}{N}(x_i - \mu) + 2(x_i - \mu) & i = j \\ -\frac{2}{N}(x_i - \mu) & i = j \end{cases}$
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 $i \neq j$:

$$\frac{\partial \hat{x}_i}{\partial x_j} = -\frac{1}{N}(v + \epsilon)^{-\frac{1}{2}} + (x_i - \mu)(-\frac{1}{2})(v + \epsilon)^{-\frac{3}{2}}\frac{2}{N}(x_j - \mu) = \\ -\frac{1}{N}(v + \epsilon)^{-\frac{1}{2}}(1 + (x_i - \mu)(x_j - \mu)(v + \epsilon)^{-1})$$

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 $\frac{\partial f}{\partial x_j}$:

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

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