Advanced Data Analysis
Homework Week 7

Aswin Vijay

June 12, 2023

1] Math

We are asked to derive $\frac{\partial J}{\partial x_i}$ where x_i is the $i'th$ units pre-activation value and J is the loss. We are given $\frac{\partial J}{\partial u_i}$, where u_i is the output after the batch normalization operation. We also have,

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i \quad mini - batchmean \tag{1}$$

$$\sigma^2_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_\beta)^2 \quad mini - batch\ variance \tag{2}$$

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \quad normalized\ x_i \tag{3}$$

$$u_i = \gamma \hat{x}_i + \beta \quad Scale\ and\ shift \tag{4}$$

To derive $\frac{\partial J}{\partial x_i}$ we write out the partial derivatives using chain rule since $\hat{x}_i(x_i, \mu, \sigma^2)$, $\sigma^2(x_i, \mu)$ and $\mu(x_i)$, three variables are dependent on x_i so:

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

The above derivatives are computed as follows,

$$\frac{\partial J}{\partial \hat{x}_i} = \frac{\partial J}{\partial u_i} \cdot \frac{\partial u_i}{\partial \hat{x}_i} = \gamma \cdot \frac{\partial J}{\partial u_i}\ From\ (4)$$

$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma^2 + \epsilon}}\ From\ (3)$$

Computing J derivatives w.r.t μ,

$$\frac{\partial J}{\partial \mu} = \sum_{i=1}^{m} \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu} + \frac{\partial J}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu} \; Summing\ over\ batch$$

$$\frac{\partial \hat{x}_i}{\partial \mu} = \frac{-1}{\sqrt{\sigma^2 + \epsilon}} \; From\ (3)$$

$$\frac{\partial \sigma^2}{\partial \mu} = \frac{-2}{m} \sum_{i=1}^{m} (x_i - \mu) = -2\left(\frac{1}{m} \sum_{i=1}^{m} x_i - \mu \right) = 0 \; From\ (2,1)$$

Computing J derivatives w.r.t σ,

$$\frac{\partial J}{\partial \sigma^2} = \sum_{i=1}^{m} \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma^2} \; Summing\ over\ batch$$

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

Putting everything together we have,

$$\frac{\partial J}{\partial \mu} = \sum_{i=1}^{m} \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2 + \epsilon}} + 0, \; then$$

$$\frac{\partial J}{\partial x_i} = \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2 + \epsilon}} + \sum_{i=1}^{m} \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2 + \epsilon}} \cdot \frac{\partial \mu}{\partial x_i} + \frac{-1}{2} (\sigma^2 + \epsilon)^{-\frac{3}{2}} \sum_{j=1}^{m} \frac{\partial J}{\partial \hat{x}_j} \cdot (x_j - \mu) \cdot \frac{\partial \sigma^2}{\partial x_i}$$

The remaining derivatives are calculated as below,

$$\frac{\partial \mu}{\partial x_i} = \frac{1}{m} \; From\ (1)$$

$$\frac{\partial \sigma^2}{\partial x_i} = \frac{2(x_i - \mu)}{m} \; From\ (2)$$

Finally we get,

$$\begin{aligned}
\frac{\partial J}{\partial x_i} &= \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2 + \epsilon}} + \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2 + \epsilon}} - \frac{(\sigma^2 + \epsilon)^{-\frac{3}{2}}}{m} \sum_{j=1}^{m} \frac{\partial J}{\partial \hat{x}_j} \cdot (x_j - \mu) \cdot (x_i - \mu) \\
&= \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2 + \epsilon}} + \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2 + \epsilon}} - \frac{1}{m\sqrt{\sigma^2 + \epsilon}} \cdot \hat{x}_i \cdot \sum_{j=1}^{m} \frac{\partial J}{\partial \hat{x}_j} \hat{x}_j \\
&= \frac{1}{m\sqrt{\sigma^2 + \epsilon}} \left(m \frac{\partial J}{\partial \hat{x}_i} - \sum_{i=1}^{m} \frac{\partial J}{\partial \hat{x}_i} - \hat{x}_i \sum_{j=1}^{m} \frac{\partial J}{\partial \hat{x}_j} \hat{x}_j \right)
\end{aligned}$$

2] Architecture

In Convolutional Neural Networks the bias parameter is associated with each filter and its value is added to the filter output. Such a bias is called a tied

bias where the bias remains constant for each location of a feature map, here the learnable parameters is lesser. Untied biases can also be used in which case each location on the input map can have its own bias, the number of learnable parameters drastically increases but the network would now be able to learn location specific information allowing for more fine tuning. Their use depends on the training data, if the data is shifted uniformly over all features a tied bias would work well, but if it has location specific shifts an untied bias can be used for correction.

If there is batch normalization operation after the convolution then biasing would be point less as the bias would also get normalized in the case of tied biases. The bias term thus becomes redundant. In case of untied biases the effect will remain as each neuron gets biased differently.