Batch Normalization equations

Arjun Majumdar

January 11, 2021

1 Cancellation of *bias* term

According to Andrew Ng's lecture on Batch Normalization, it is mentioned that: "What BatchNorm (BN) does is that it looks at a mini-batch and normalizes z^l to first have mean 0 and unit standard variance and then re-scales z^l by using β and γ. But what that means is that whatever is the value of the bias b^l, it just gets subtracted out because during the BN normalization step, you compute the mean of z^l and subtract out the mean. So, adding any constant b^l to all of the examples in a mini-batch does not change anything. Because any constant (b^l) that you add will get canceled out by the mean subtraction step."

$$z^{[l]} = W^{[l]} \cdot a^{[l-1]} + b^{[l]} \qquad \text{network input for layer 'l'} \tag{1.1}$$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} z^{[l](i)} \qquad \text{mean for layer 'l'} \tag{1.2}$$

$$= W^{[l]} \cdot \overline{a[l-1]} + b[l] \tag{1.3}$$

$$= W^{[l]} \cdot \frac{1}{m}\sum_{i=1}^{m} a^{[l-1](i)} + b^{[l]} \tag{1.4}$$

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} (z^{[l](i)} - \mu)^2 \qquad \text{variance for layer 'l'} \tag{1.5}$$

$$z^{[l]}_{norm} = \frac{z^{[l]} - \mu}{\sqrt{\sigma^2 + \epsilon}} \qquad \text{normalization for layer 'l'} \tag{1.6}$$

$$\tilde{z^{[l]}} = \gamma^{[l]} \cdot z^{[l]}_{norm} + \beta^{[l]} \qquad \text{batch normalized input for layer 'l'} \tag{1.7}$$

In the equations above, 'm' is number of training examples.
$\mu = W\bar{a} + b$ is the term μ containing 'b'.

In order to mathematically show that subtracting the mean leads to cancellation of the bias term (or, $b^{[l]}$), we have:

$$
\begin{aligned}
z^{[l]} - \mu &= (W^{[l]} \cdot a^{[l-1]} + b^{[l]}) - \mu && (1.8) \\
&= (W^{[l]} \cdot a^{[l-1]} + b^{[l]}) - \left(W^{[l]} \cdot \frac{1}{m} \sum_{i=1}^{m} a^{[l-1](i)} + b^{[l]} \right) && (1.9) \\
&= W^{[l]} \cdot a^{[l-1]} + b^{[l]} - W^{[l]} \cdot \frac{1}{m} \sum_{i=1}^{m} a^{[l-1](i)} - b^{[l]} && (1.10) \\
&= W^{[l]} \cdot a^{[l-1]} + \cancel{b^{[l]}} - W^{[l]} \cdot \frac{1}{m} \sum_{i=1}^{m} a^{[l-1](i)} - \cancel{b^{[l]}} && (1.11) \\
&= W^{[l]} \cdot \left(a^{[l-1]} - \frac{1}{m} \sum_{i=1}^{m} a^{[l-1](i)} \right) && (1.12)
\end{aligned}
$$

Hence, the bias term (or, $b^{[l]}$) is canceled and we can replace equation 1.6 with equation 1.12.