Batch Normalization

Normalizing inputs to speed up learning

Given some intermediate values in NN, unit i for layer l

given as $Z^{[l](i)} = \{ Z^{(1)}, Z^{(2)}, \ldots, Z^{(m)} \}$

where each values are from some inputs of a mini batch.

so $\mu = \frac{1}{m} \sum_i Z^{(i)}$ $\quad \sigma^2 = \frac{1}{m} \sum_i (Z^{(i)} - \mu)^2$

$Z^{(i)}_{norm} = \frac{Z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$

$\tilde{Z}^{(i)} = \gamma Z^{(i)}_{norm} + \beta.$

γ & β are learnable.

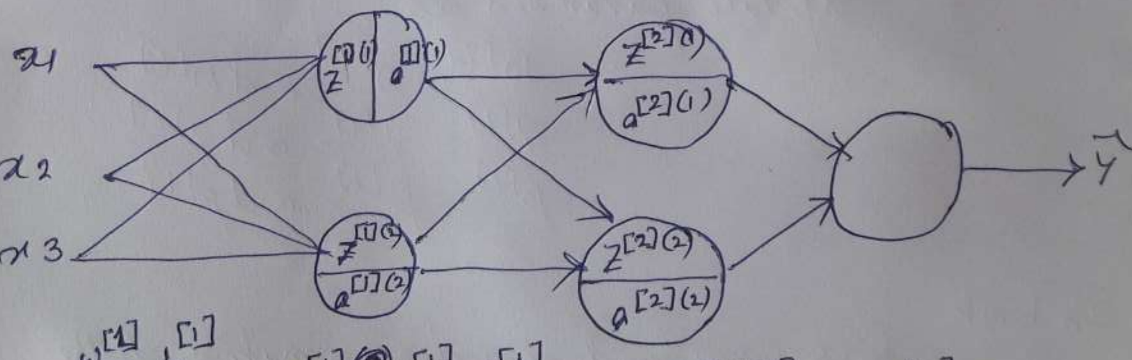
If $\gamma = 1$, $\beta = 0$

$\tilde{Z}^{(i)} = Z^{(i)}_{norm}.$

If $\gamma = \sqrt{\sigma^2 + \epsilon}$ $\beta = \mu$

then $\tilde{Z}^{(i)} = Z^{(i)}.$

Filling Batch norm in a neural network



$X \xrightarrow{W^{[1]}, b^{[1]}} Z^{[1]} \xrightarrow[\text{Batch norm}]{\beta^{[1]} \ \gamma^{[1]}} \tilde{Z}^{[1]} \longrightarrow a^{[1]} = g^{[1]}(\tilde{Z}^{[1]})$

$\downarrow W^{[2]}, b^{[2]}$

$\cdots \longleftarrow a^{[2]} \longleftarrow \tilde{Z}^{[2]} \xleftarrow[\text{Batch norm}]{\beta^{[2]}, \gamma^{[2]}} Z^{[2]}$

Parameters $W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]} \ldots$

$\gamma^{[1]}, \beta^{[1]}, \gamma^{[2]}, \beta^{[2]} \ldots$

The bias term may be removed from calculation

as $Z^{[L]} = W^{[L]} a^{[L-1]} + b^{[L]}$

so $\mu^{[L](i)} = \frac{1}{m} \sum_{i=1}^{m} Z^{[L](i)} = P^{(i)} + b^{[L](i)}$

The bias term remains in μ

so in $Z^{(i)} - \mu$, the bias term gets removed.

Implementing Gradient descent with BN

for $k = 1$ to m (# of minibatches)

do forward propagation using $X^{(k)}$

In each hidden layer use BN to replace $Z^{[L]}$ with $\tilde{Z}^{[L]}$

Use back prop to compute $dW^{[L]}$, $d\gamma^{[L]}$ & $d\beta^{[L]}$

Update parameters as

$$W^{[L]} = W^{[L]} - \alpha\, dW^{[L]}$$
$$\gamma^{[L]} = \gamma^{[L]} - \alpha\, d\gamma^{[L]}$$
$$\beta^{[L]} = \beta^{[L]} - \alpha\, d\beta^{[L]}$$

In book

H is considered as a design matrix of the activation of a minibatch

$$H = X^{(i)} \downarrow i \quad \xrightarrow{\text{minibatch } k} \quad = \begin{array}{c|c|c} X^{\{1\}(1)} & X^{\{2\}(1)} & X^{\{3\}(1)} \\ \hline X^{\{1\}(2)} & X^{\{2\}(2)} & X^{\{3\}(2)} \end{array}$$

$$\mu = \begin{bmatrix} \text{mean} \\ \text{of} \\ \text{rows} \end{bmatrix} \quad \sigma = \begin{bmatrix} \text{variance} \\ \text{of} \\ \text{rows} \end{bmatrix}$$

$$H' = \frac{H - \mu}{\sigma}$$