Batch Normalisation - Implementation from Scratch

Christian Meo, Francisco Castanheira and Marco Sala
Delft University of Technology

April 2020

1 Network Architecture - Batch Normalization

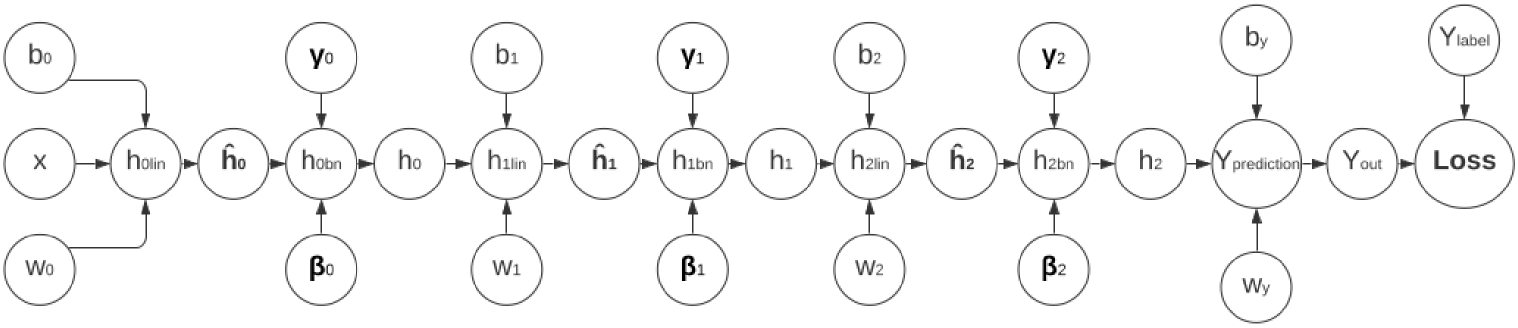

Figure 1: Scheme of the Architecture

2 Dimensions

Dimensions of the different parameters, for a better comprehension. N is the batch size, D_{in} is the input dimension, H is the hidden dimension and D_{out} is the output dimension.

$$\begin{array}{ccc} x & \rightarrow & N \times D_{in} \\ w_0 & \rightarrow & D_{in} \times H \\ b_0, b_1, b_2 & \rightarrow & 1 \times H \\ w_1, w_2 & \rightarrow & H \times H \\ w_y & \rightarrow & H \times D_{out} \\ b_y & \rightarrow & 1 \times D_{out} \\ \gamma_0, \gamma_1, \gamma_2, & \rightarrow & 1 \times H \\ \beta_0, \beta_1, \beta_2, & \rightarrow & 1 \times H \\ y_{label} & \rightarrow & N \times 1 \end{array} \tag{1}$$

3 Forward pass

While $\epsilon = 10^{-8}$ and `ones` represents a vectors of ones with dimensions N x 1. This vector was used as an intuitive way to understand summations over the batch dimension, for instance. Additionally, $\odot$ represents elementwise multiplication.

$$h_{0lin} = w_0 \cdot x + b_0 \rightarrow N \times H \tag{2}$$

This will not be mentioned again, but in cases like the one above, where the vector being summed or multiplied element wise is one-dimensional, it has to be broadcast with the dimension remaining dimension of the matrix as an input.

$$\mu_0 = E[h_{0lin}] \quad \text{along the batch dimension} \ \rightarrow 1 \times H \tag{3}$$

$$\sigma_0^2 = E[(h_{0lin} - \mu_0)^2] \quad \text{along the batch dimension } \rightarrow 1 \times H \tag{4}$$

$$\hat{h_0} = \frac{h_{0lin} - \mu_0}{\sqrt{\sigma_0^2 + \epsilon}} \rightarrow N \times H \tag{5}$$

$$h_{0BN} = \gamma_0 \odot \hat{h_0} + \beta_0 \rightarrow N \times H \tag{6}$$

$$h_0 = \texttt{sigmoid}(h_{0BN}) \rightarrow N \times H \tag{7}$$

$$h_{1lin} = w_1 \cdot h_0 + b_1 \rightarrow N \times H \tag{8}$$

$$\mu_1 = E[h_{1lin}] \quad \text{along the batch dimension } \rightarrow 1 \times H \tag{9}$$

$$\sigma_1^2 = E[(h_{1lin} - \mu_1)^2] \quad \text{along the batch dimension } \rightarrow 1 \times H \tag{10}$$

$$\hat{h_1} = \frac{h_{1lin} - \mu_1}{\sqrt{\sigma_1^2 + \epsilon}} \rightarrow N \times H \tag{11}$$

$$h_{1BN} = \gamma_1 \cdot \hat{h_1} + \beta_1 \rightarrow N \times H \tag{12}$$

$$h_1 = \texttt{sigmoid}(h_{1BN}) \rightarrow N \times H \tag{13}$$

$$h_{2lin} = w_2 \cdot h_1 + b_2 \rightarrow N \times H \tag{14}$$

$$\mu_2 = E[h_{2lin}] \quad \text{along the batch dimension } \rightarrow 1 \times H \tag{15}$$

$$\sigma_2^2 = E[(h_{2lin} - \mu_2)^2] \quad \text{along the batch dimension } \rightarrow 1 \times H \tag{16}$$

$$\hat{h_2} = \frac{h_{2lin} - \mu_1}{\sqrt{\sigma_2^2 + \epsilon}} \rightarrow N \times H \tag{17}$$

$$h_{2BN} = \gamma \cdot \hat{h_2} + \beta \rightarrow N \times H \tag{18}$$

$$h_2 = \texttt{sigmoid}(h_{2BN}) \rightarrow N \times H \tag{19}$$

$$y_{pred} = w_y \cdot h_2 + b_y \rightarrow N \times D_{out} \tag{20}$$

$$y_{out} = \texttt{softmax}(y_{pred}) \rightarrow N \times D_{out} \tag{21}$$

$$L = \texttt{cross\_entropy}(y_{out}, y_{label}) \rightarrow N \times D_{out} \tag{22}$$

4 Back prop

$$\frac{\partial L}{\partial y_{pred}} = \frac{\partial L}{\partial y_{out}}\frac{\partial y_{out}}{\partial y_{pred}} = y_{out} - (y_{label})_{\text{onehot}} \rightarrow N \times D_{out} \tag{23}$$

$$\frac{\partial L}{\partial w_y} = \frac{\partial L}{\partial y_{pred}}\frac{\partial y_{pred}}{\partial w_y} = h_2^T \cdot \frac{\partial L}{\partial y_{pred}} \rightarrow H \times D_{out} \tag{24}$$

$$\frac{\partial L}{\partial b_y} = \frac{\partial L}{\partial y_{pred}}\frac{\partial y_{pred}}{\partial b_y} = (\texttt{ones})^T \frac{\partial L}{\partial y_{pred}} \rightarrow 1 \times D_{out} \tag{25}$$

$$\frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial y_{pred}}\frac{\partial y_{pred}}{\partial h_2} = \frac{\partial L}{\partial y_{pred}} \cdot w_y^T \rightarrow N \times H \tag{26}$$

$$\frac{\partial L}{\partial h_{2BN}} = \frac{\partial L}{\partial h_2}\frac{\partial h_2}{\partial h_{2BN}} = \frac{\partial L}{\partial h_2} \odot [h_2 \odot (1 - h_2)] \rightarrow N \times H \tag{27}$$

$$\frac{\partial L}{\partial \gamma_2} = \frac{\partial L}{\partial h_{2BN}}\frac{\partial h_{2BN}}{\partial \gamma_2} = (\texttt{ones})^T \cdot \left(\frac{\partial L}{\partial h_{2BN}} \odot \hat{h_2}\right) \rightarrow 1 \times H \tag{28}$$

$$\frac{\partial L}{\partial \beta_2} = \frac{\partial L}{\partial h_{2BN}}\frac{\partial h_{2BN}}{\partial \beta_2} = (\texttt{ones})^T \cdot \frac{\partial L}{\partial h_{2BN}} \rightarrow 1 \times H \tag{29}$$

$$\frac{\partial L}{\partial \hat{h_2}} = \frac{\partial L}{\partial h_{2BN}}\frac{\partial h_{2BN}}{\partial \hat{h_2}} = \frac{\partial L}{\partial h_{2BN}} \odot \gamma_2 \rightarrow N \times H \tag{30}$$

$$\frac{\partial L}{\partial \sigma_2^2} = (\texttt{ones})^T \cdot \left[\frac{\partial L}{\partial \hat{h_2}} \odot (h_{2lin} - \mu_2) \odot (\sigma_2^2 + \epsilon)^{\frac{-3}{2}}\right] \cdot \left(-\frac{1}{2}\right) \rightarrow 1 \times H \tag{31}$$

$$\frac{\partial L}{\partial \mu_2} = (\texttt{ones})^T \cdot \left(\frac{\partial L}{\partial \hat{h_2}} \odot \frac{-1}{\sqrt{\sigma_2^2 + \epsilon}}\right) + \frac{\partial L}{\partial \sigma_2^2} \odot \left((\texttt{ones})^T \cdot \frac{-2 \cdot (h_{2lin} - \mu_2)}{N}\right) \rightarrow 1 \times H \tag{32}$$

$$\frac{\partial L}{\partial h_{2lin}} = \frac{\partial L}{\partial \hat{h_2}} \odot \frac{1}{\sqrt{\sigma_2^2 + \epsilon}} + \frac{\partial L}{\partial \sigma_2^2} \odot \frac{2 \cdot (h_{2lin} - \mu_2)}{N} + \frac{\partial L}{\partial \mu_2} \cdot \frac{1}{N} \rightarrow N \times H \tag{33}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial h_{2lin}}\frac{\partial h_{2lin}}{\partial w_2} = h_1^T \cdot \frac{\partial L}{\partial h_{2lin}} \rightarrow H \times H \tag{34}$$

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial h_{2lin}}\frac{\partial h_{2lin}}{\partial b_2} = (\texttt{ones})^T \cdot \frac{\partial L}{\partial h_{2lin}} \rightarrow 1 \times H \tag{35}$$

$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial h_{2lin}}\frac{\partial h_{2lin}}{\partial h_1} = \frac{\partial L}{\partial h_{2lin}} \cdot w_2^T \rightarrow N \times H \tag{36}$$

$$\frac{\partial L}{\partial h_{1BN}} = \frac{\partial L}{\partial h_1}\frac{\partial h_1}{\partial h_{1BN}} = \frac{\partial L}{\partial h_1} \odot [h_1 \odot (1 - h_1)] \rightarrow N \times H \tag{37}$$

$$\frac{\partial L}{\partial \gamma_1} = \frac{\partial L}{\partial h_{1BN}}\frac{\partial h_{1BN}}{\partial \gamma_1} = (\texttt{ones})^T \cdot \left(\frac{\partial L}{\partial h_{1BN}} \odot \hat{h_1}\right) \rightarrow 1 \times H \tag{38}$$

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial L}{\partial h_{1BN}}\frac{\partial h_{1BN}}{\partial \beta_1} = (\texttt{ones})^T \cdot \frac{\partial L}{\partial h_{1BN}} \rightarrow 1 \times H \tag{39}$$

$$\frac{\partial L}{\partial \hat{h_1}} = \frac{\partial L}{\partial h_{1BN}}\frac{\partial h_{1BN}}{\partial \hat{h_1}} = \frac{\partial L}{\partial h_{1BN}} \odot \gamma_1 \rightarrow N \times H \tag{40}$$

$$\frac{\partial L}{\partial \sigma_1^2} = (\texttt{ones})^T \cdot \left[\frac{\partial L}{\partial \hat{h_1}} \odot (h_{1lin} - \mu_1) \odot (\sigma_1^2 + \epsilon)^{\frac{-3}{2}}\right] \cdot \left(-\frac{1}{2}\right) \rightarrow 1 \times H \tag{41}$$

$$\frac{\partial L}{\partial \mu_1} = (\texttt{ones})^T \cdot \left(\frac{\partial L}{\partial \hat{h_1}} \odot \frac{-1}{\sqrt{\sigma_1^2 + \epsilon}}\right) + \frac{\partial L}{\partial \sigma_1^2} \odot \left((\texttt{ones})^T \cdot \frac{-2 \cdot (h_{1lin} - \mu_1)}{N}\right) \rightarrow 1 \times H \tag{42}$$

$$\frac{\partial L}{\partial h_{1lin}} = \frac{\partial L}{\partial \hat{h_1}} \odot \frac{1}{\sqrt{\sigma_1^2 + \epsilon}} + \frac{\partial L}{\partial \sigma_1^2} \odot \frac{2 \cdot (h_{1lin} - \mu_1)}{N} + \frac{\partial L}{\partial \mu_1} \cdot \frac{1}{N} \rightarrow N \times H \tag{43}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial h_{1lin}} \frac{\partial h_{1lin}}{\partial w_1} = h_0^T \cdot \frac{\partial L}{\partial h_{1lin}} \rightarrow H \times H \tag{44}$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial h_{1lin}} \frac{\partial h_{1lin}}{\partial b_1} = (\texttt{ones})^T \cdot \frac{\partial L}{\partial h_{1lin}} \rightarrow 1 \times H \tag{45}$$

$$\frac{\partial L}{\partial h_0} = \frac{\partial L}{\partial h_{1lin}} \frac{\partial h_{1lin}}{\partial h_0} = \frac{\partial L}{\partial h_{1lin}} \cdot w_1^T \rightarrow N \times H \tag{46}$$

$$\frac{\partial L}{\partial h_{0BN}} = \frac{\partial L}{\partial h_0} \frac{\partial h_0}{\partial h_{0BN}} = \frac{\partial L}{\partial h_0} \odot [h_0 \odot (1 - h_0)] \rightarrow N \times H \tag{47}$$

$$\frac{\partial L}{\partial \gamma_0} = \frac{\partial L}{\partial h_{0BN}} \frac{\partial h_{0BN}}{\partial \gamma_0} = (\texttt{ones})^T \cdot \left(\frac{\partial L}{\partial h_{0BN}} \odot \hat{h_0} \right) \rightarrow 1 \times H \tag{48}$$

$$\frac{\partial L}{\partial \beta_0} = \frac{\partial L}{\partial h_{0BN}} \frac{\partial h_{0BN}}{\partial \beta_0} = (\texttt{ones})^T \cdot \frac{\partial L}{\partial h_{0BN}} \rightarrow 1 \times H \tag{49}$$

$$\frac{\partial L}{\partial \hat{h_0}} = \frac{\partial L}{\partial h_{0BN}} \frac{\partial h_{0BN}}{\partial \hat{h_0}} = \frac{\partial L}{\partial h_{0BN}} \odot \gamma_0 \rightarrow N \times H \tag{50}$$

$$\frac{\partial L}{\partial \sigma_0^2} = (\texttt{ones})^T \cdot \left[\frac{\partial L}{\partial \hat{h_0}} \odot (h_{0lin} - \mu_0) \odot (\sigma_0^2 + \epsilon)^{\frac{-3}{2}} \right] \cdot \left(-\frac{1}{2} \right) \rightarrow 1 \times H \tag{51}$$

$$\frac{\partial L}{\partial \mu_0} = (\texttt{ones})^T \cdot \left(\frac{\partial L}{\partial \hat{h_0}} \odot \frac{-1}{\sqrt{\sigma_0^2 + \epsilon}} \right) + \frac{\partial L}{\partial \sigma_0^2} \odot \left((\texttt{ones})^T \cdot \frac{-2 \cdot (h_{0lin} - \mu_0)}{N} \right) \rightarrow 1 \times H \tag{52}$$

$$\frac{\partial L}{\partial h_{0lin}} = \frac{\partial L}{\partial \hat{h_0}} \odot \frac{1}{\sqrt{\sigma_0^2 + \epsilon}} + \frac{\partial L}{\partial \sigma_0^2} \odot \frac{2 \cdot (h_{0lin} - \mu_0)}{N} + \frac{\partial L}{\partial \mu_0} \cdot \frac{1}{N} \rightarrow N \times H \tag{53}$$

$$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial h_{0lin}} \frac{\partial h_{0lin}}{\partial w_0} = x^T \cdot \frac{\partial L}{\partial h_{0lin}} \rightarrow D_{in} \times H \tag{54}$$

$$\frac{\partial L}{\partial b_0} = \frac{\partial L}{\partial h_{0lin}} \frac{\partial h_{0lin}}{\partial b_0} = (\texttt{ones})^T \cdot \frac{\partial L}{\partial h_{0lin}} \rightarrow 1 \times H \tag{55}$$