Back propagation in BatchNorm

Author: Aritra Roy Gosthipaty Date: 12 August 2020

Feed Forward:

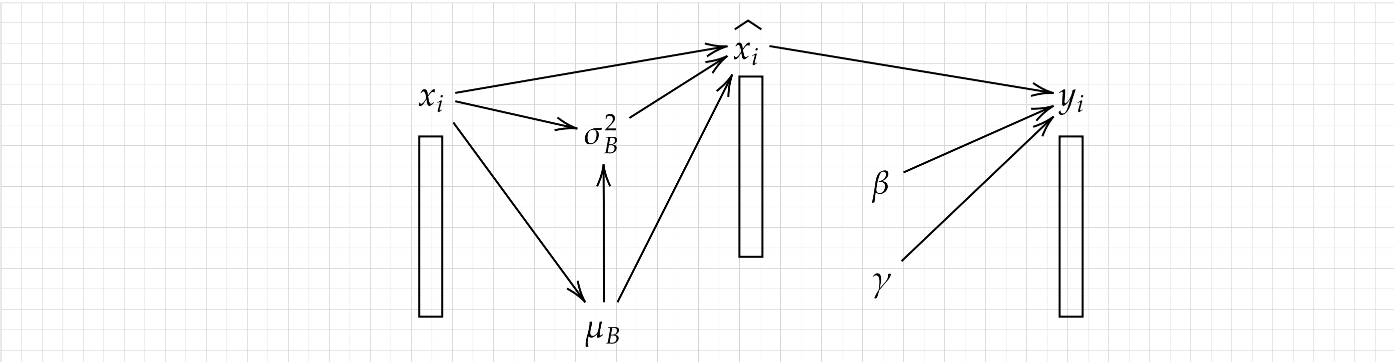
We consider a mini-batch B, of size m. The i^{th} element in any one dimension of activation is represented as the x_i. Actually we consider x_i^k as the k^{th} dim and i^{th} element, but to keep things concise, the k is taken out of the derivation. The mean and variance of the mini-batch are $\mu_B \; and \; \sigma_B^2$ respectively. $\gamma \; and \; \beta$ are the scaling and shifting parameters of the batch-norm layer.

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2 \tag{2}$$

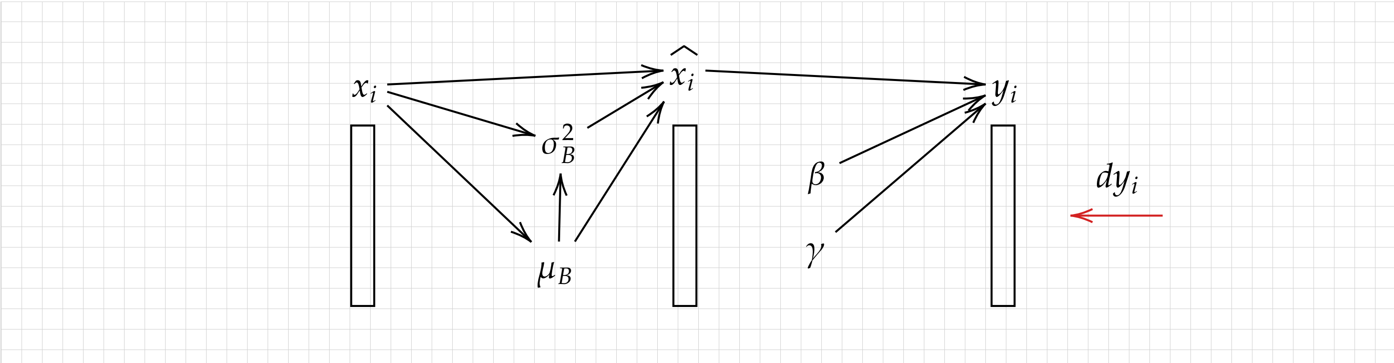
$$\widehat{x_i} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{3}$$

$$y_i = \gamma \widehat{x_i} + \beta \tag{4}$$



Back Propagation:

Let us consider that we have $\frac{\partial l}{\partial y_i}$ flowing upstream into our network. We will back-prop into every parameter in the batch-norm with the help of chain rule. For our convenience we will replace $\frac{\partial l}{\partial a}$ where a is any parameter, with da.

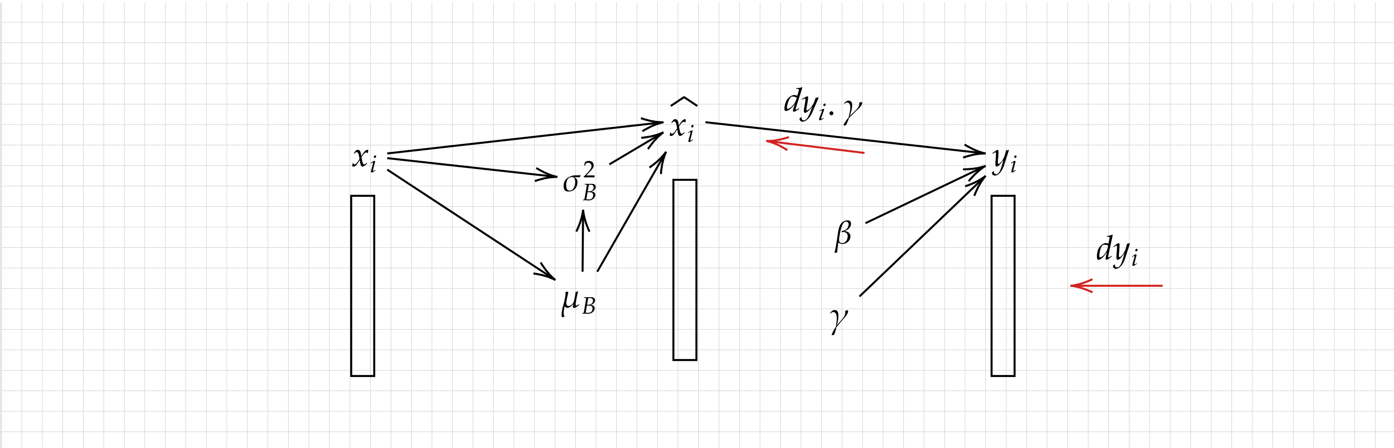


$$Diff\ (4)\ wrt\ \widehat{x_i}\ we\ get$$

$$\frac{\partial y_i}{\partial \widehat{x_i}} = \gamma \tag{5}$$

$$\frac{\partial l}{\partial \widehat{x_i}} = \frac{\partial l}{\partial y_i} . \frac{\partial y_i}{\partial \widehat{x_i}}$$

$$\implies \frac{\partial l}{\partial \widehat{x_i}} = dy_i . \gamma \tag{From 5}$$



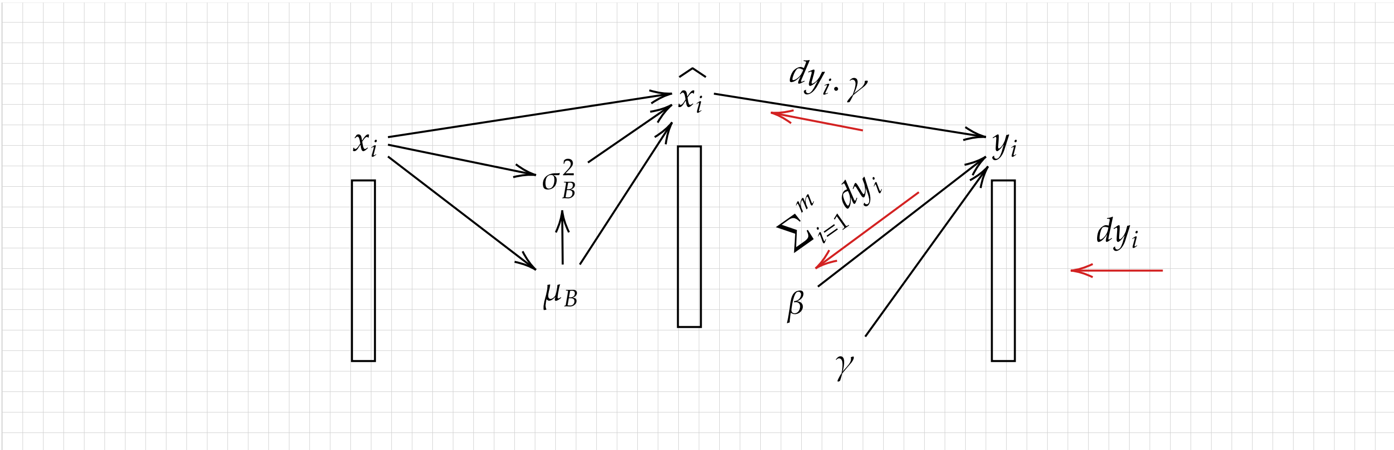
Note to the reader: When the gradient dy_i flows into the network, each of the i^{th} element of $\widehat{x_i}$ is effected by the corresponding i^{th} element of dy_i. Now to consider all the collective gradient flow for the single valued $\beta\ and\ \gamma$ we need to add the gradients flowing in.

$$Diff\ (4)\ wrt\ \beta\ we\ get$$

$$\frac{\partial y_i}{\partial \beta} = 1 \tag{6}$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} . \frac{\partial y_i}{\partial \beta}$$

$$\implies \frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} dy_i \tag{From 6}$$

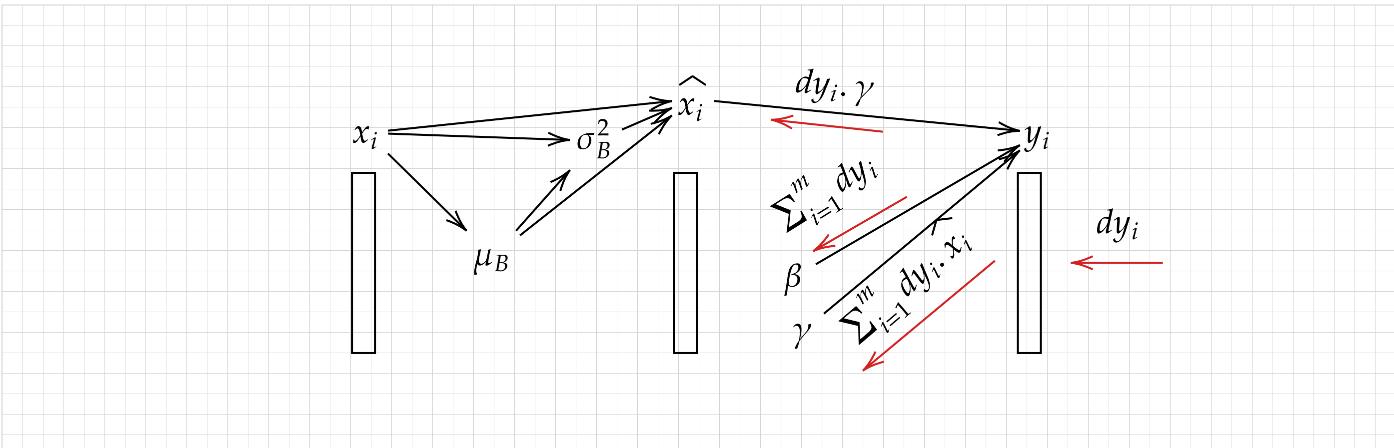


$$Diff\ (4)\ wrt\ \gamma\ we\ get$$

$$\frac{\partial y_i}{\partial \gamma} = \widehat{x_i} \tag{7}$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} . \frac{\partial y_i}{\partial \gamma}$$

$$\implies \frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} dy_i . \widehat{x_i} \qquad \text{(From 7)}$$

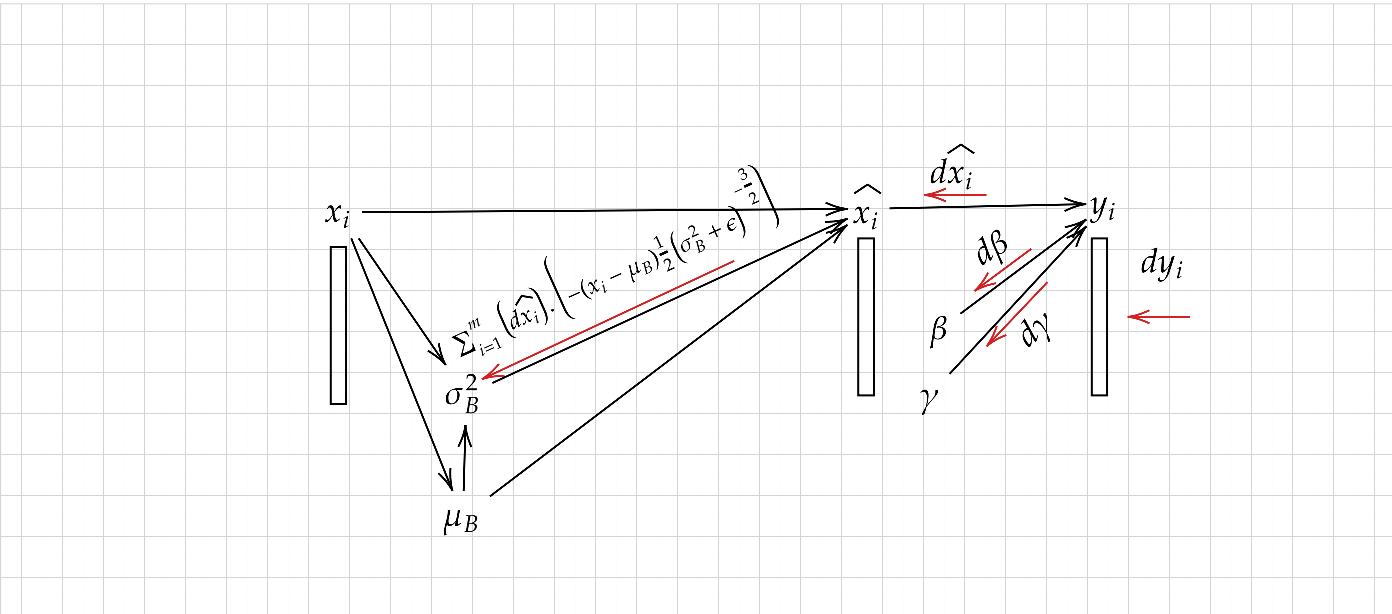


A note for the reader: When the gradient $\widehat{dx_i}$ flows into the network, each of the i^{th} element of x_i is effected by the corresponding i^{th} element of $\widehat{dx_i}$. Now to consider all the collective gradient flow for single valued $\mu_B\ and\ \sigma_B^2$ we need to add the gradients flowing in.

$$Diff\ (3)\ wrt\ \sigma_B^2$$

$$
\begin{aligned}
\frac{\partial \widehat{x_i}}{\partial \sigma_B^2} &= \frac{\left(\sqrt{\sigma_B^2+\epsilon}\right)(0) - (x_i-\mu_B)\frac{1}{2}\left(\sigma_B^2+\epsilon\right)^{-\frac{1}{2}}}{\sigma_B^2+\epsilon} \\
\implies \frac{\partial \widehat{x_i}}{\partial \sigma_B^2} &= -(x_i-\mu_B)\frac{1}{2}\left(\sigma_B^2+\epsilon\right)^{-\frac{1}{2}-1} \\
\implies \frac{\partial \widehat{x_i}}{\partial \sigma_B^2} &= -(x_i-\mu_B)\frac{1}{2}\left(\sigma_B^2+\epsilon\right)^{-\frac{3}{2}}
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma_B^2} &= \sum_{i=1}^{m} \frac{\partial l}{\partial y_i}.\frac{\partial y_i}{\partial \widehat{x_i}}.\frac{\partial \widehat{x_i}}{\partial \sigma_B^2} \\
\implies \frac{\partial l}{\partial \sigma_B^2} &= \sum_{i=1}^{m} \frac{\partial l}{\partial \widehat{x_i}}.\frac{\partial \widehat{x_i}}{\partial \sigma_B^2} \\
\implies \frac{\partial l}{\partial \sigma_B^2} &= \sum_{i=1}^{m} d\widehat{x_i}.\frac{\partial \widehat{x_i}}{\partial \sigma_B^2} \\
\implies \frac{\partial l}{\partial \sigma_B^2} &= \sum_{i=1}^{m} \left(d\widehat{x_i}\right).\left(-(x_i-\mu_B)\frac{1}{2}\left(\sigma_B^2+\epsilon\right)^{-\frac{3}{2}}\right) && \text{(From 8)}
\end{aligned}
$$



$$Diff\ (2)\ wrt\ \mu_B$$

$$
\frac{\partial \sigma_B^2}{\partial \mu_B} = \frac{\partial\left(\frac{1}{m}\Sigma_{i=1}^{m}(x_i-\mu_B)^2\right)}{\partial \mu_B}
$$

$$\implies \frac{\partial \sigma_B^2}{\partial \mu_B} = \frac{1}{m}\sum_{i=1}^{m} -2(x_i - \mu_B) \tag{9}$$

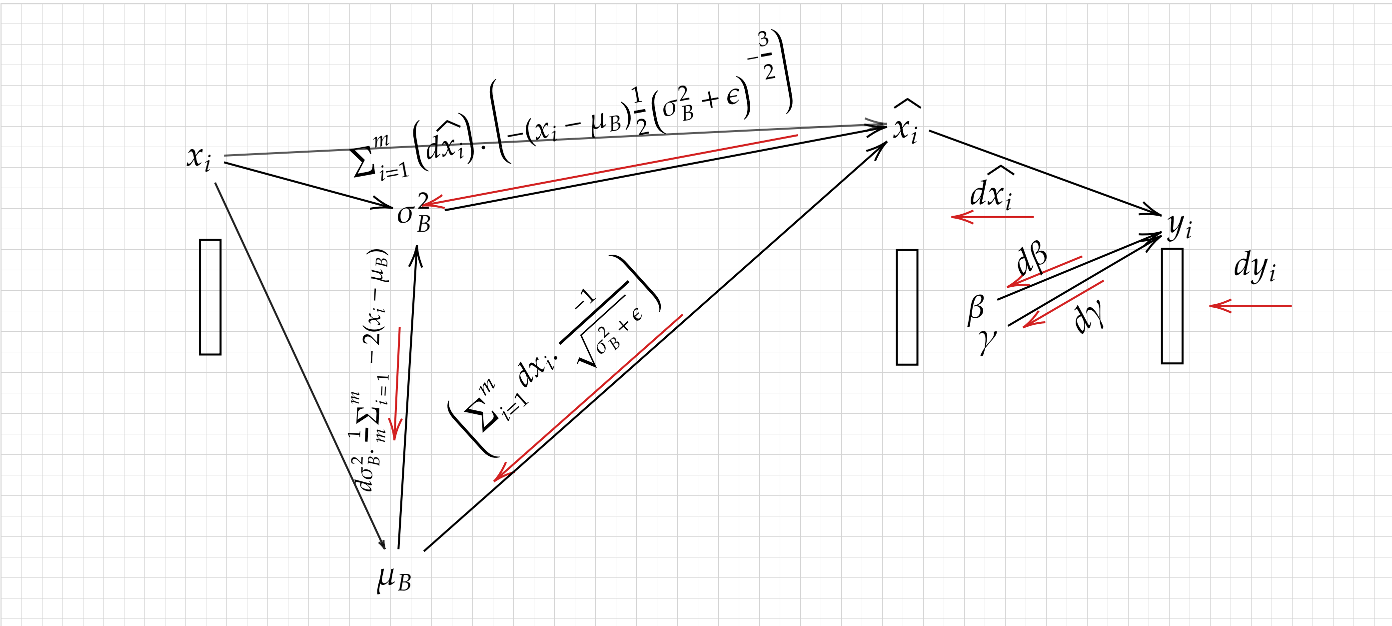
$$Diff\ (3)\ wrt\ \mu_B$$

$$\frac{\partial \widehat{x_i}}{\partial \mu_B} = \frac{\partial\left(\frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}\right)}{\partial \mu_B}$$

$$\implies \frac{\partial \widehat{x_i}}{\partial \mu_B} = \frac{-1}{\sqrt{\sigma_B^2 + \epsilon}} \tag{10}$$

$$\frac{\partial l}{\partial \mu_B} = \left(\sum_{i=1}^{m} \frac{\partial l}{\partial \widehat{x_i}} . \frac{\partial \widehat{x_i}}{\partial \mu_B}\right) + \frac{\partial l}{\partial \sigma_B^2} . \frac{\partial \sigma_B^2}{\partial \mu_B}$$

$$\implies \frac{\partial l}{\partial \mu_B} = \left(\sum_{i=1}^{m} dx_i . \frac{-1}{\sqrt{\sigma_B^2 + \epsilon}}\right) + d\sigma_B^2 . \frac{1}{m}\sum_{i=1}^{m} -2(x_i - \mu_B) \qquad \text{(From 9 \& 10)}$$



$$Diff\ (1)\ wrt\ x_i,$$

removing the summation sign as the grad is done element wise

$$\implies \frac{\partial \mu_B}{\partial x_i} = \frac{\partial\left(\frac{1}{m}x_i\right)}{\partial x_i}$$

$$\implies \frac{\partial \mu_B}{\partial x_i} = \frac{1}{m} \tag{11}$$

$$Diff\ (2)\ wrt\ x_i$$

$$\frac{\partial \sigma_B^2}{\partial x_i} = \frac{\partial\left(\frac{1}{m}(x_i-\mu_B)^2\right)}{\partial x_i}$$

$$\implies \frac{\partial \sigma_B^2}{\partial x_i} = \frac{1}{m}2(x_i-\mu_B) \tag{12}$$

$$Diff\ (3)\ wrt\ x_i$$

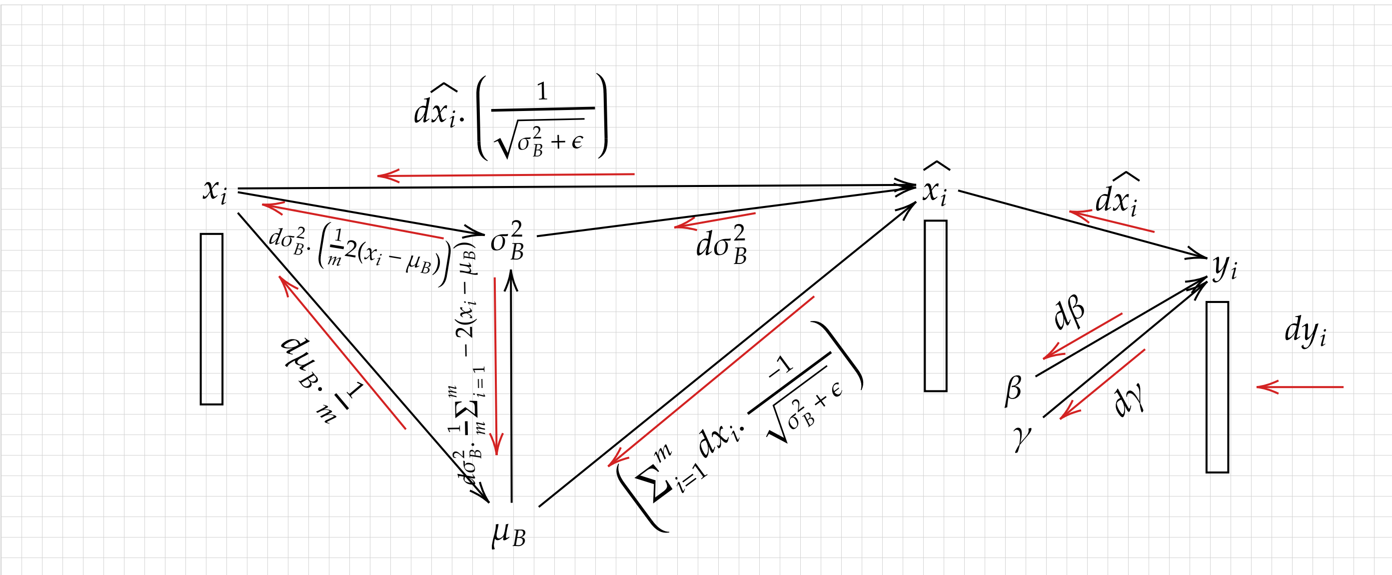
$$\frac{\partial \widehat{x_i}}{\partial x_i} = \frac{\partial\left(\frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\epsilon}}\right)}{\partial x_i}$$

$$\implies \frac{\partial \widehat{x_i}}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2+\epsilon}} \tag{13}$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \widehat{x_i}}.\frac{\partial \widehat{x_i}}{\partial x_i} + \frac{\partial l}{\partial \sigma_B^2}.\frac{\partial \sigma_B^2}{\partial x_i} + \frac{\partial l}{\partial \mu_B}.\frac{\partial \mu_B}{\partial x_i}$$

$$From\ (11),\ (12)\ and\ (13)$$

$$\implies \frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \widehat{x_i}}.\left(\frac{1}{\sqrt{\sigma_B^2+\epsilon}}\right) + \frac{\partial l}{\partial \sigma_B^2}.\left(\frac{1}{m}2(x_i-\mu_B)\right) + \frac{\partial l}{\partial \mu_B}.\frac{1}{m}$$



Conclusion:

Batch-Normalization is one of the best innovations that have been done in this field. I was unable to figure out the layer's derivative sheme that is suggested, so started from scratch. This is not the usual blog that I write, instead this is completely written in latex and hence is in the pdf format. I ask the reader to pardon my poor skills in latex. I hope the reader had a good time going through the derivatives.