Back propagation in BatchNorm

Author: Aritra Roy Gosthipaty Date: 12 August 2020

Introduction:

I was reading the paper on BatchNorm [Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift; by Sergey Ioffe and Christian Szegedy] and stumbled upon a number of equations.

$$
\begin{aligned}
\frac{\partial \ell}{\partial \widehat{x}_i} &= \frac{\partial \ell}{\partial y_i} \cdot \gamma \\
\frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \widehat{x}_i} \cdot (x_i - \mu_{\mathcal{B}}) \cdot \frac{-1}{2}(\sigma_{\mathcal{B}}^2 + \epsilon)^{-3/2} \\
\frac{\partial \ell}{\partial \mu_{\mathcal{B}}} &= \left(\sum_{i=1}^{m} \frac{\partial \ell}{\partial \widehat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \right) + \frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} \cdot \frac{\sum_{i=1}^{m} -2(x_i - \mu_{\mathcal{B}})}{m} \\
\frac{\partial \ell}{\partial x_i} &= \frac{\partial \ell}{\partial \widehat{x}_i} \cdot \frac{1}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} \cdot \frac{2(x_i - \mu_{\mathcal{B}})}{m} + \frac{\partial \ell}{\partial \mu_{\mathcal{B}}} \cdot \frac{1}{m} \\
\frac{\partial \ell}{\partial \gamma} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \widehat{x}_i \\
\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i}
\end{aligned}
$$

Figure 1: The back propagation through the batch norm layer

These equations are responsible for the backward propagation through a batch norm layer. Even after reading the equations multiple times I found the equations very unintuitive. This led me to sit down with my notepad and scribble the forward and backward propagation graphs. I thought of uploading a scan of my notepad but that would not have been helpful at all (my handwriting can kill people, in the negative sense). Here I am providing my sketches and derivations to make sense of what the authors say through the equations.

Aside: It would really help if you open the paper along side this blog post. The notations used here are exactly the same as that of the paper.

Feed Forward:

An excerpt from the paper will familiarize the reader to the notations used.

"Consider a min-atch B of size m. Since the normalization is applied to each activation independently, let us focus on a particular activation $x^{(k)}$ and omit k for clarity. We have m values of activation in the mini-batch;

$$B = \{x_{1...m}\}$$

Let the normalized values be $\widehat{x}_{1...m}$, and their linear transformations be $y_{1...m}$."

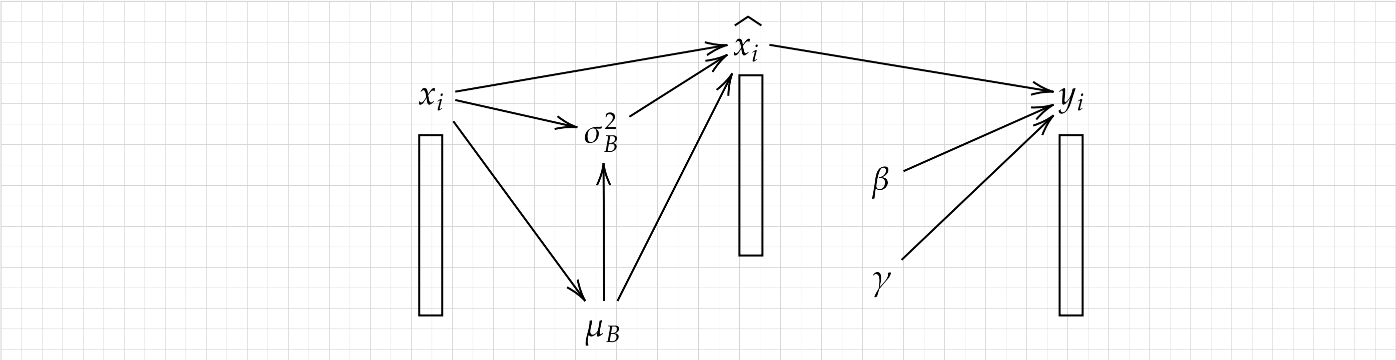
The mean and variance of the mini-batch are μ_B *and* σ_B^2 respectively. γ *and* β are the scaling and shifting parameters of the batch-norm layer.

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

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

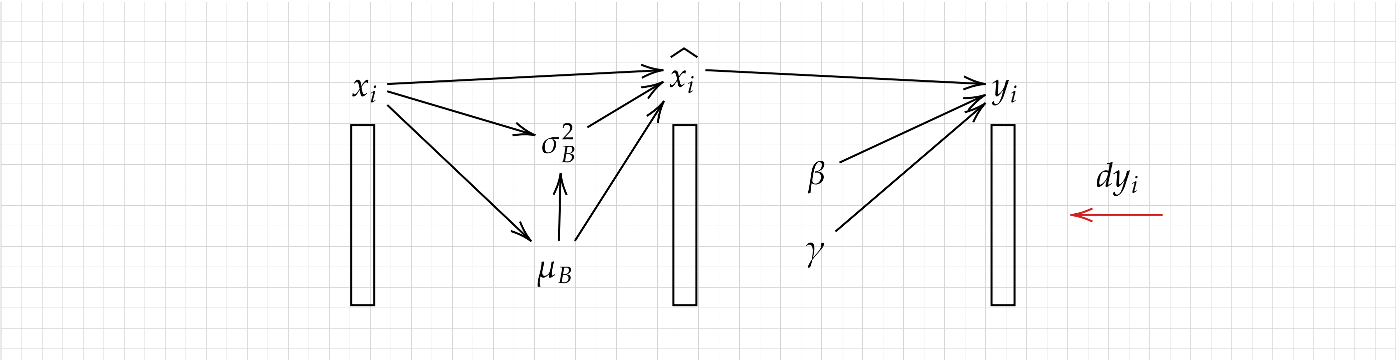
$$\widehat{x_i} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{3}$$

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



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.

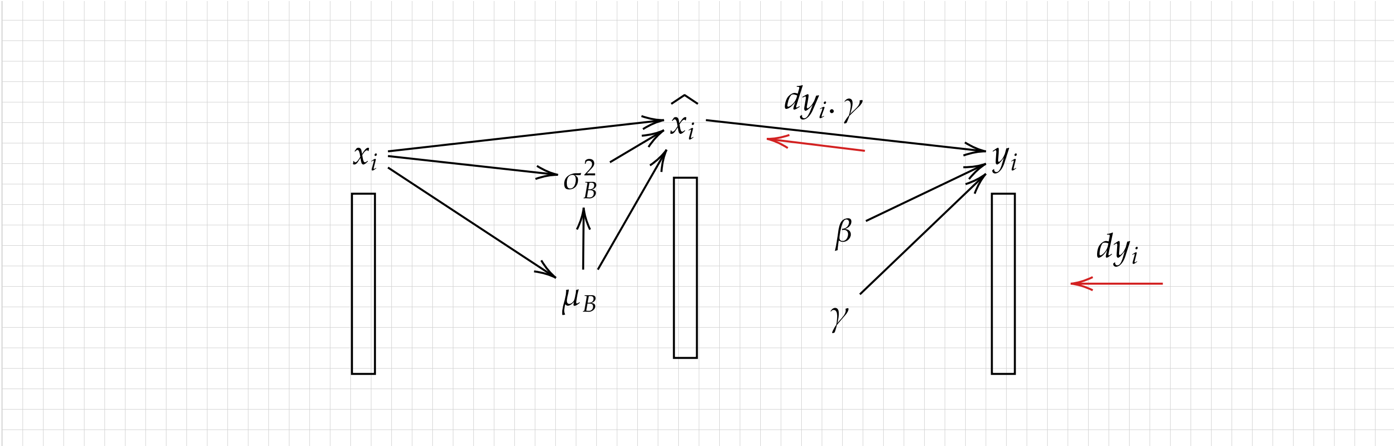


$$Diff\ (4)\ wrt\ \widehat{x_i}\ we\ get$$

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

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

$$\implies \frac{\partial l}{\partial \widehat{x_i}} = dy_i.\gamma \quad \text{(From 5)}$$



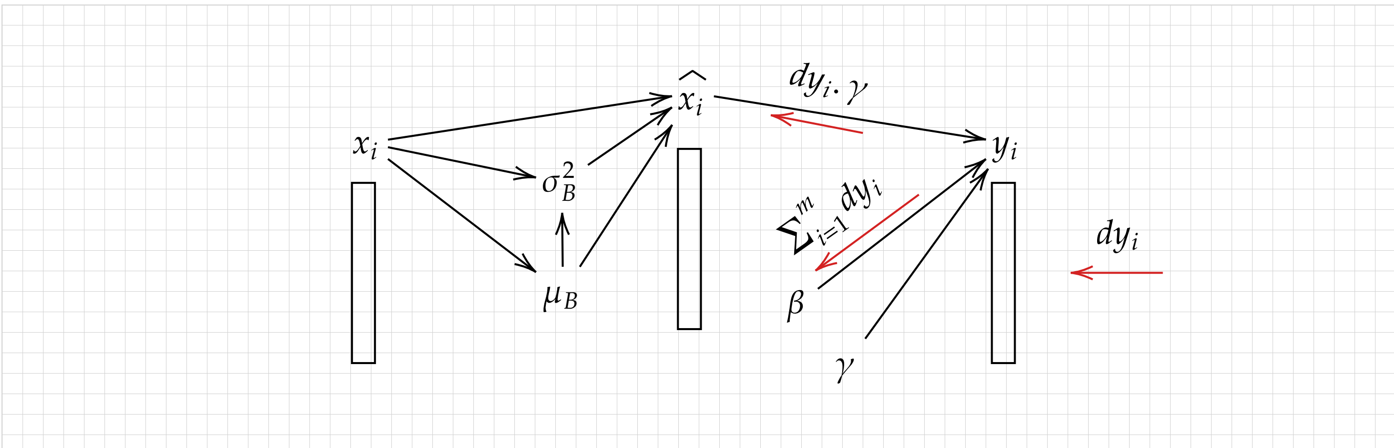
Note to the reader: When the gradient dy_i flows into the network, each of the i^{th} element of $\widehat{x_i}$ is effected by the corresponding i^{th} element of dy_i. Now to consider all the collective gradient flow for the single valued $\beta\ and\ \gamma$ we need to *add* the gradients flowing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

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

$$\frac{\partial y_i}{\partial \beta} = 1 \quad (6)$$

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

$$\implies \frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} dy_i \quad \text{(From 6)}$$

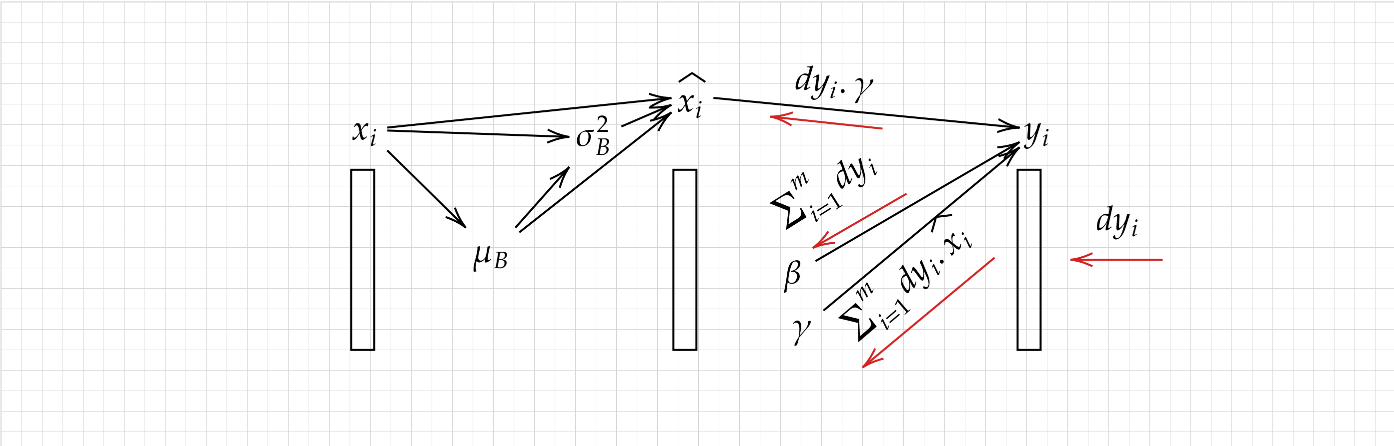


$$Diff\ (4)\ wrt\ \gamma\ we\ get$$

$$\frac{\partial y_i}{\partial \gamma} = \widehat{x_i} \quad (7)$$

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

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



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 μ_B and σ_B^2 we need to add the gradients flowing in.

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

$$\frac{\partial \widehat{x_i}}{\partial \sigma_B^2} = \frac{\left(\sqrt{\sigma_B^2 + \epsilon}\right)(0) - (x_i - \mu_B)\frac{1}{2}\left(\sigma_B^2 + \epsilon\right)^{-\frac{1}{2}}}{\sigma_B^2 + \epsilon}$$

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

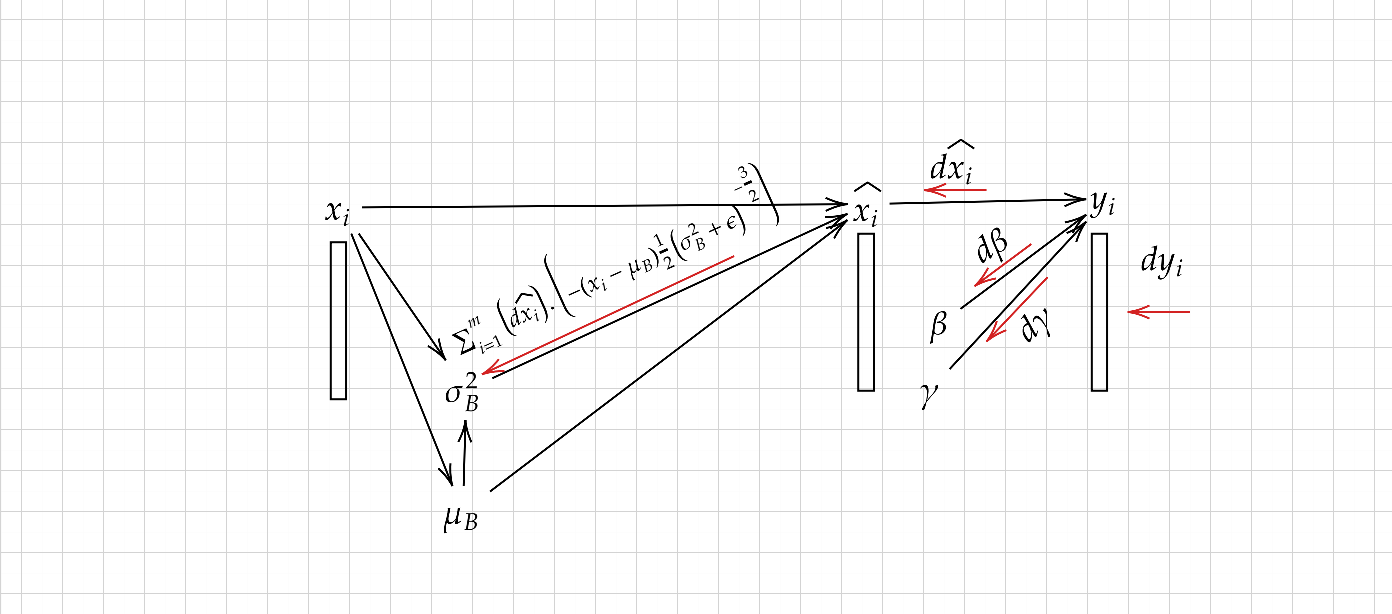
$$\implies \frac{\partial \widehat{x_i}}{\partial \sigma_B^2} = -(x_i - \mu_B)\frac{1}{2}\left(\sigma_B^2 + \epsilon\right)^{-\frac{3}{2}} \qquad (8)$$

$$\frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} . \frac{\partial y_i}{\partial \widehat{x_i}} . \frac{\partial \widehat{x_i}}{\partial \sigma_B^2}$$

$$\implies \frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial l}{\partial \widehat{x_i}} . \frac{\partial \widehat{x_i}}{\partial \sigma_B^2}$$

$$\implies \frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^{m} \widehat{dx_i} . \frac{\partial \widehat{x_i}}{\partial \sigma_B^2}$$

$$\implies \frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^{m} \left(\widehat{dx_i}\right).\left(-(x_i-\mu_B)\frac{1}{2}\left(\sigma_B^2+\epsilon\right)^{-\frac{3}{2}}\right) \qquad \text{(From 8)}$$



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

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

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

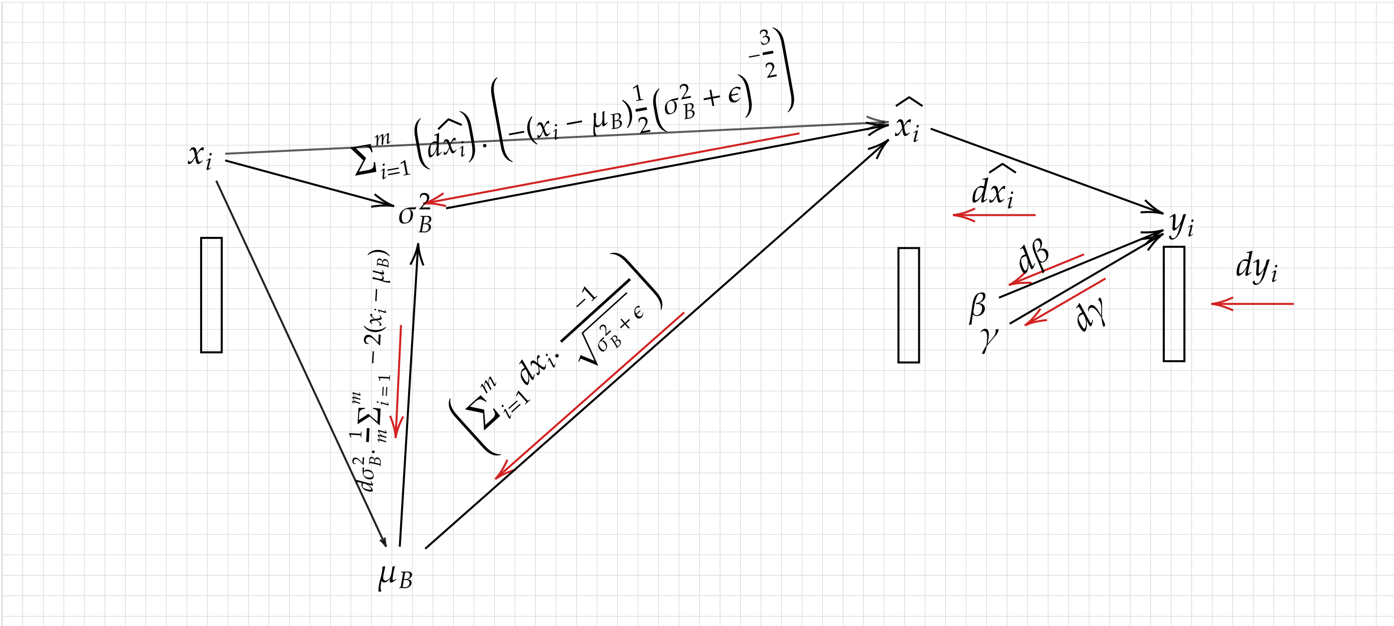
$$Diff\ (3)\ wrt\ \mu_B$$

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

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

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

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



$$Diff\ (1)\ wrt\ x_i,$$
$$removing\ the\ summation\ sign\ as\ the\ grad\ is\ done\ element\ wise$$

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

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

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

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

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

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

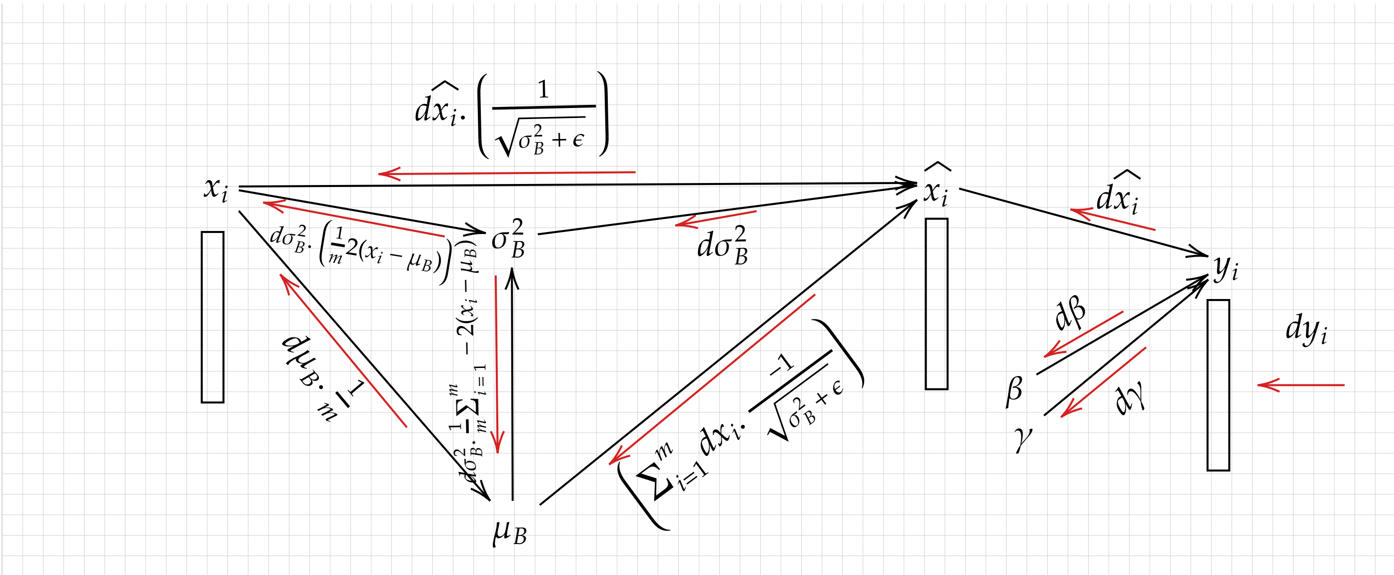
$$\frac{\partial \widehat{x_i}}{\partial x_i} = \frac{\partial\left(\frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\epsilon}}\right)}{\partial x_i}$$

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

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

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

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



Final thoughts:

Pardon my poor drawing and LaTeX skills. This is not a blog per say but a piece that helps build the intuitions for the process. I hope the reader is clear with the process and can visualize how beautiful the idea of batch norm is.

I would love to hear from the reader on any discrepancies and extensions to this work. Thank you for your time.