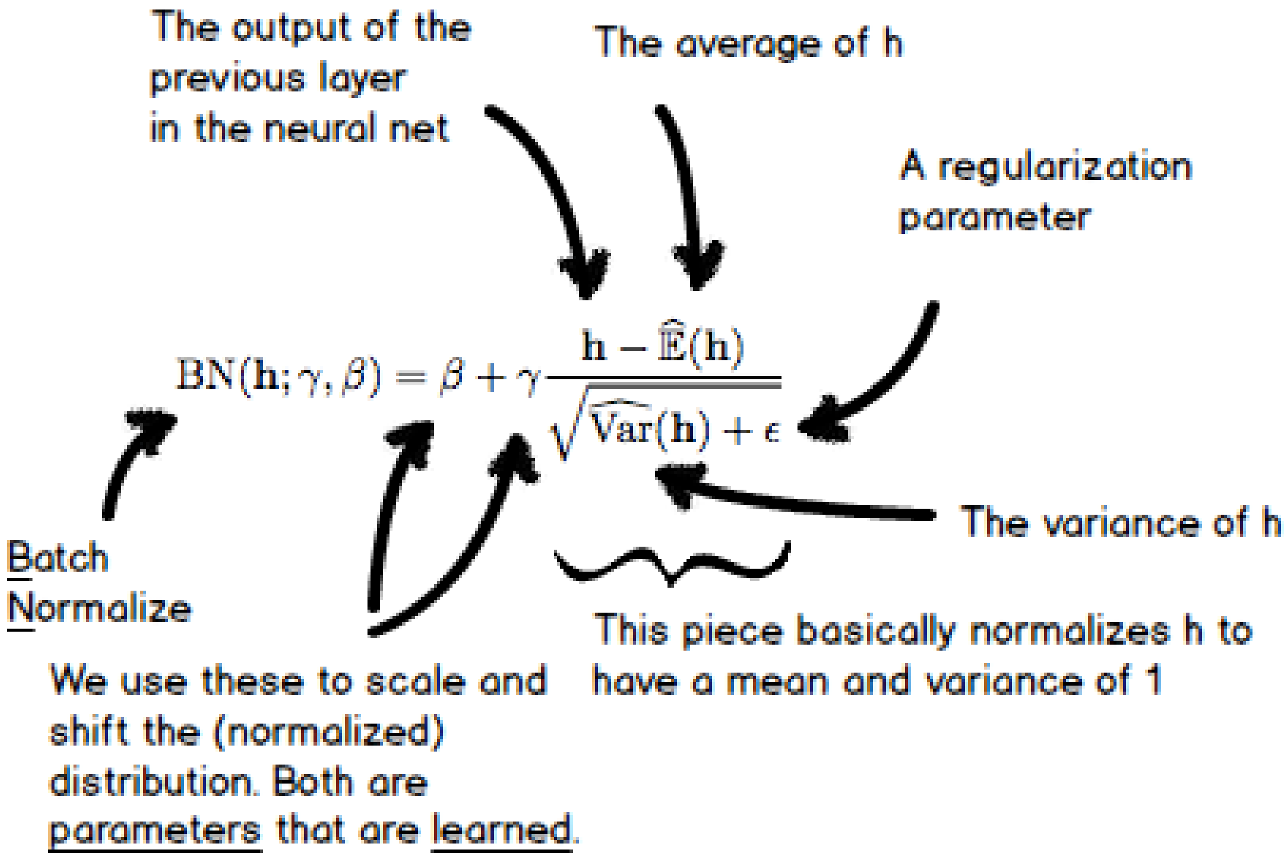
批次归一化 (Batch Normalization)

- 通常插入在卷积和全连接之后，在非线性处理前。　　位置：卷积 → BN → ReLU
- 为使每一维成为标准高斯分布(均值为0，方差为1)，可应用 $\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$
- 为能工作在激活的非线性区，再进行缩放和移位（scale&shift）处理



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

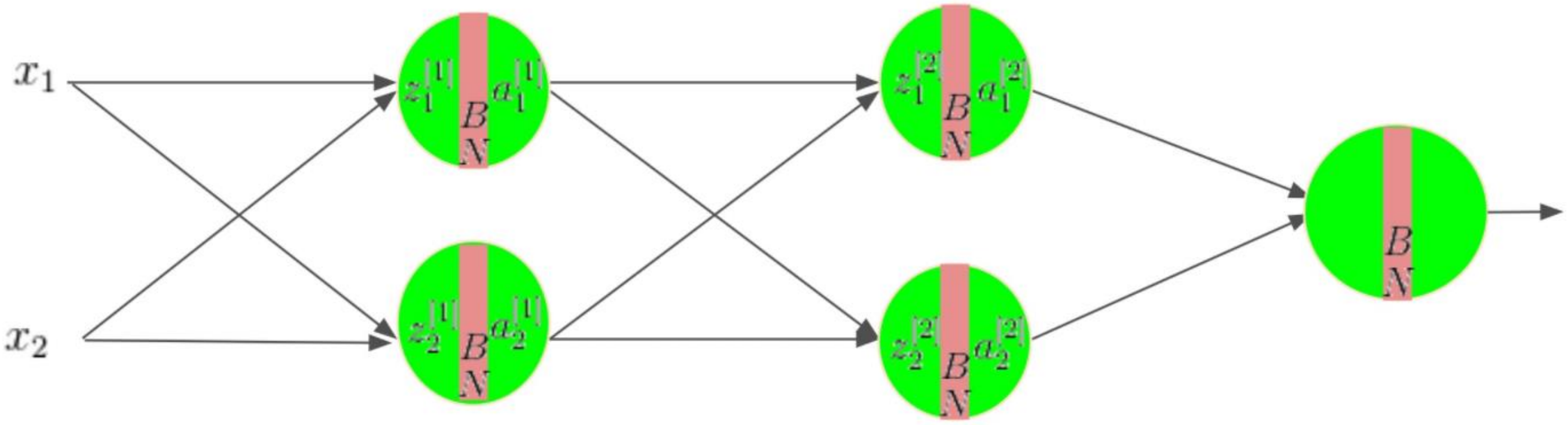
$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

在每个mini-batch中计算得到mini-batch的mean和variance来替代整体训练集的mean和variance



$$z_i^l = \left(w_i^l\right)^T a_i^{l-1} \qquad \text{(BN 1.1)}$$

$$\mu_B^l = \frac{1}{m}\sum_{i=1}^{m} z_i^l \qquad \text{(BN 1.2)}$$

$$\left(\sigma_B^2\right)^{(l)} = \frac{1}{m}\sum_{i=1}^{m}\left(z_i^l - \mu_B^l\right)^2 \qquad \text{(BN 1.3)}$$

$$\hat{z}_i^l = \frac{z_i^l - \mu_B^l}{\sqrt{\left(\sigma_B^2\right)^{(l)} + \varepsilon}} \qquad \text{(BN 1.4)}$$

$$y_i^l = \gamma \hat{z}_i^l + \beta \qquad \text{(BN 1.5)}$$

$$a_i^l = g\left(y_i^l\right) \qquad \text{(BN 1.6)}$$

预测时，计算总体的均值和方差是不实际的，也是无法实现的，因为无法采样到所有样本。

在训练过程中的batch下的均值μ_B和方差σ_B，可以加以利用来估计总体

$$\mu = E(x) = E\left(\mu_B\right) = \frac{1}{K}\sum_B^K \mu_B$$

$$D(x) = \frac{m}{m-1} E\left(\sigma_A^2\right)$$

m为batch_size

虽然理论上是如此, Darknet里面用滚动平均值来算。

$\overline{x}_m$ 表示前m个数据的平均值

滚动平均：

$$
\begin{aligned}
\overline{x}_m &= \frac{x_1 + x_2 + \cdots + x_m}{m} \\
&= \frac{x_1 + x_2 + \cdots + x_{m-1} + x_m}{m} \\
&= \frac{(x_1 + x_2 + \cdots + x_{m-1})(m-1)}{m} + \frac{x_m}{m} \\
&= \frac{m-1}{m}\overline{x}_{m-1} + \frac{x_m}{m} \\
&= \left(1 - \frac{1}{m}\right)\overline{x}_{m-1} + \frac{1}{m}x_m
\end{aligned}
$$

BN层反向传播过程

设最终的损失函数为C, 对方差求导:

$$\frac{\partial C}{\partial\left(\sigma_B^2\right)^l} = \sum_{i=1}^{m}\left(\frac{\partial C}{\partial \hat{z}_i^l}\frac{\partial \hat{z}_i^l}{\partial\left(\sigma_B^2\right)^l}\right)$$

$$= \sum_{i=1}^{m}\left\{\frac{-1}{2}\frac{\partial C}{\partial z_i^{l+1}}\frac{\partial z_i^{l+1}}{\partial a_i^l}\frac{\partial a_i^l}{\partial y_i^l}\frac{\partial y_i^l}{\partial \hat{z}_i^l}\left(z_i^l-\mu_B^l\right)\left[\left(\sigma_B^2\right)^{(l)}+\varepsilon\right]^{-\frac{3}{2}}\right\}$$

$$= \sum_{i=1}^{m}\left\{\frac{-1}{2}\left[\delta^{l+1}\frac{\partial z_i^{l+1}}{\partial a_i^l}\odot\left(g\left(y_i^l\right)'\cdot\gamma_i^l\right)\right]\left(z_i^l-\mu_B^l\right)\left[\left(\sigma_B^2\right)^{(l)}+\varepsilon\right]^{-\frac{3}{2}}\right\}$$

$$= \gamma^l\odot\sum_{i=1}^{m}\left\{\frac{-1}{2}\left[\delta^{l+1}\frac{\partial z_i^{l+1}}{\partial a_i^l}\odot\sigma\left(y_i^l\right)'\right]\left(z_i^l-\mu_B^l\right)\left[\left(\sigma_B^2\right)^{(l)}+\varepsilon\right]^{-\frac{3}{2}}\right\} \qquad \text{(BN 2.1)}$$

$$\frac{\partial C}{\partial \mu_B^l}=\sum_{i=1}^{m}\left(\frac{\partial C}{\partial \hat{z}_i}\frac{\hat{z}_i}{\partial \mu_B^l}+\frac{\partial C}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial \mu_B^l}\right)\qquad=\sum_{i=1}^{m}\left(\frac{\partial C}{\partial \hat{z}_i}\frac{-1}{\sqrt{\left(\sigma_B^2\right)^l+\varepsilon}}\right)+\frac{\partial C}{\partial\left(\sigma_B^2\right)^l}\cdot\frac{-2}{m}\cdot\sum_{i}^{m}\left(z_i^l-\mu_B^l\right)$$

$$\sum_{i=1}^{m}\left(\frac{\partial C}{\partial \hat{z}_i}\frac{-1}{\sqrt{\left(\sigma_B^2\right)^l+\varepsilon}}\right)\qquad=\sum_{i=1}^{m}\left(\frac{\partial C}{\partial z_i^{l+1}}\frac{\partial z_i^{l+1}}{\partial a_i^l}\frac{\partial a_i^l}{\partial y_i^l}\frac{\partial y_i^l}{\partial \hat{z}_i^l}\frac{-1}{\sqrt{\left(\sigma_B^2\right)^l+\varepsilon}}\right)$$

$$=\gamma^l\odot\sum_{i=1}^{m}\left(\left[\delta^{l+1}\frac{\partial z_i^{l+1}}{\partial a_i^l}\odot g\left(y_i^l\right)'\right]\frac{-1}{\sqrt{\left(\sigma_B^2\right)^l+\varepsilon}}\right)\qquad\text{(BN 2.2)}$$

$$\delta_i^l = \frac{\partial C}{\partial z_i^l} = \frac{\partial C}{\partial \hat{z}_i^l}\frac{\partial \hat{z}_i^l}{\partial z_i^l} + \frac{\partial C}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial z_i^l} + \frac{\partial C}{\partial \mu_B}\frac{\partial \mu_B}{\partial z_i^l}$$

$$= \frac{\partial C}{\partial \hat{z}_i^l}\frac{\mathbf{1}}{\sqrt{\sigma_B^2 + \varepsilon}} + \frac{\partial C}{\partial \sigma_B^2}\cdot\frac{\mathbf{2}}{m}\cdot\left(z_i^l - \mu_B\right) + \frac{\partial C}{\partial \mu_B}\cdot\frac{\mathbf{1}}{m}$$

$$= \frac{\partial C}{\partial z_i^{l+\mathbf{1}}}\frac{\partial z_i^{l+\mathbf{1}}}{\partial a_i^l}\frac{\partial a_i^l}{\partial y_i^l}\frac{\partial y_i^l}{\partial \hat{z}_i^l} + \frac{\partial C}{\sqrt{\sigma_B^2 + \varepsilon}} + \frac{\partial C}{\partial \sigma_B^2}\cdot\frac{\mathbf{2}}{m}\cdot\left(z_i^l - \mu_B\right) + \frac{\partial C}{\partial \mu_B}\cdot\frac{\mathbf{1}}{m}$$

$$= \left[\delta^{l+\mathbf{1}}\frac{\partial z_i^{l+\mathbf{1}}}{\partial a_i^l} \odot \left(g\left(y_i^l\right)' \odot \gamma_i^l\right)\right]\frac{\mathbf{1}}{\sqrt{\sigma_B^2 + \varepsilon}} + \frac{\partial C}{\partial \sigma_B^2}\cdot\frac{\mathbf{2}}{m}\cdot\left(z_i^l - \mu_B\right) + \frac{\partial C}{\partial \mu_B}\cdot\frac{\mathbf{1}}{m} \qquad \text{(BN 2.3)}$$

求权重和偏差梯度：

$$\frac{\partial C}{\partial w^l} = \sum_{i=1}^{m} \frac{\partial C}{\partial z_i^l} \frac{\partial z_i^l}{\partial w^l} = \sum_{i=1}^{m} \left(a^{l-1}\right)^T \delta^l \qquad \text{(BN 2.4)}$$

$$\frac{\partial C}{\partial \beta^l} = \sum_{i=1}^{m} \frac{\partial C}{\partial z_i^{l+1}} \frac{\partial z_i^{l+1}}{\partial a_i^l} \frac{\partial a_i^l}{\partial y_i^l} \frac{\partial y_i^l}{\partial \beta^l} = \sum_{i=1}^{m} \delta_i^{l+1} \frac{\partial z_i^{l+1}}{\partial a_i^l} \odot \left[g\left(y_i^l\right)' \frac{\partial y^l}{\partial \beta^l} \right]$$

$$= \sum_{i=1}^{m} \delta_i^{l+1} \frac{\partial z_i^{l+1}}{\partial a_i^L} \odot g\left(y_i^l\right)' \qquad \text{(BN 2.5)}$$

$$\frac{\partial C}{\partial \gamma^l} = \sum_{i=1}^{m} \frac{\partial C}{\partial z_i^{l+1}} \frac{\partial z_i^{l+1}}{\partial a_i^l} \frac{\partial a_i^l}{\partial y_i^l} \frac{\partial y_i^l}{\partial \gamma^l} = \sum_{i=1}^{m} \delta_i^{l+1} \frac{\partial z_i^{l+1}}{\partial a_i^l} \odot \left[g\left(y_i^l\right)' \odot \hat{z}_i^l \right] \qquad \text{(BN 2.6)}$$