1. Обратное распространение ошибки из оригинальной статьи про $\texttt{Batch normalisation}\left(\texttt{https://arxiv.org/pdf/1502.03167.pdf}\right)$:

$$\frac{\partial l}{\partial \hat{x}_i} = \frac{\partial l}{\partial y_i} \cdot \gamma$$

$$\frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot (x_i - \mu_B) \cdot \left(-\frac{1}{2}\right) \left(\sigma_B^2 + \epsilon\right)^{-3/2}$$

$$\frac{\partial l}{\partial \mu_B} = \left(\sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot \left(-\frac{1}{\sqrt{\sigma_B^2 + \epsilon}}\right)\right) + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{\sum_{i=1}^{m} -2\,(x_i - \mu_B)}{m}$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{2\,(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B^2} \cdot \frac{1}{m}$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \cdot \hat{x}_i$$

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

2. Соберем наиболее важные обозначения из оригинальной статьи:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sigma_B + \epsilon} \; - \; \text{z-score для } x_i$$

$$y_i = \gamma\, x_i + \beta$$

$$\frac{\partial l}{\partial \hat{x}_i} = \frac{\partial l}{\partial y_i} \cdot \gamma$$

3. Разберем $\frac{\partial l}{\partial x_i}$ для упрощения вычслений обратного распространения ошибки:

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{2\,(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B^2} \cdot \frac{1}{m} \quad \left(\text{Развернём } \frac{\partial l}{\partial \sigma_B^2}\right)$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} + \left[\sum_{i=1}^{m} \frac{\partial l}{\partial \hat{x}_i} \cdot (x_i - \mu_B) \cdot \left(-\frac{1}{2}\right) \left(\sigma_B^2 + \epsilon\right)^{-3/2}\right] \cdot \frac{2\,(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B^2} \cdot \frac{1}{m}$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} - \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot (x_j - \mu_B) \cdot \frac{1}{\left(\sqrt{\sigma_B^2 + \epsilon}\right)^3}\right] \cdot \frac{(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B^2} \cdot \frac{1}{m}$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} - \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \frac{(x_j - \mu_B)}{\sqrt{\sigma_B^2 + \epsilon}} \cdot \frac{1}{\left(\sqrt{\sigma_B^2 + \epsilon}\right)^2}\right] \cdot \frac{(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B^2} \cdot \frac{1}{m}$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} - \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \hat{x}_j \cdot \frac{1}{\left(\sqrt{\sigma_B^2 + \epsilon}\right)^2}\right] \cdot \frac{(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B^2} \cdot \frac{1}{m}$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} - \frac{1}{m\sqrt{\sigma_B^2 + \epsilon}}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \hat{x}_j\right] \cdot \hat{x}_i + \frac{\partial l}{\partial \mu_B^2} \cdot \frac{1}{m} \quad \left(\text{Развернём } \frac{\partial l}{\partial \mu_B^2}\right)$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} - \frac{\hat{x}_i}{m\sqrt{\sigma_B^2 + \epsilon}} \cdot \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \hat{x}_j\right] +$$
$$\frac{1}{m} \cdot \left[\left(\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \cdot \left(-\frac{1}{\sqrt{\sigma_B^2 + \epsilon}}\right)\right) + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{\sum_{j=1}^{m} -2\,(x_j - \mu_B)}{m}\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m\sqrt{\sigma_B^2 + \epsilon}}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] -$$
$$\frac{1}{m\sqrt{\sigma_B^2 + \epsilon}}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right] + \frac{1}{m^2}\left[\frac{\partial l}{\partial \sigma_B^2} \sum_{j=1}^{m} -2\,(x_j - \mu_B)\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) +$$
$$\frac{1}{m^2}\left[\frac{\partial l}{\partial \sigma_B^2} \sum_{j=1}^{m} -2\,(x_j - \mu_B)\right] \quad \left(\text{Развернём } \frac{\partial l}{\partial \sigma_B^2}\right)$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m^2}\left[\frac{\partial l}{\partial \sigma_B^2} \sum_{j=1}^{m} (x_j - \mu_B)\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m}\left[\frac{\partial l}{\partial \sigma_B^2} \frac{\sum_{j=1}^{m} (x_j - \mu_B)}{m}\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m}\left[\frac{\partial l}{\partial \sigma_B^2} \left(\frac{\sum_{j=1}^{m} x_j}{m} - \frac{\sum_{j=1}^{m} \mu_B}{m}\right)\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m}\left[\frac{\partial l}{\partial \sigma_B^2} \left(\frac{\sum_{j=1}^{m} x_j}{m} - \frac{m}{m} \mu_B\right)\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m}\left[\frac{\partial l}{\partial \sigma_B^2} \left(\frac{\sum_{j=1}^{m} x_j}{m} - \mu_B\right)\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m}\left[\frac{\partial l}{\partial \sigma_B^2} (\mu_B - \mu_B)\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m}\left[\frac{\partial l}{\partial \sigma_B^2} 0\right]$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j\right] - \frac{1}{m}\left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j}\right]\right) + \frac{(-2)}{m} 0$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m} \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j \right] - \frac{1}{m} \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \right] \right) + 0$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m} \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j \right] - \frac{1}{m} \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \right] \right)$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{m}{m} \frac{\partial l}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m} \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j \right] - \frac{1}{m} \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \right] \right)$$

$$\frac{\partial l}{\partial x_i} = \frac{1}{m \sqrt{\sigma_B^2 + \epsilon}} \left(m \frac{\partial l}{\partial \hat{x}_i} - \hat{x}_i \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \hat{x}_j \right] - \left[\sum_{j=1}^{m} \frac{\partial l}{\partial \hat{x}_j} \right] \right)$$

Последнее выражение является упрощенной версией вычисления $\frac{\partial l}{\partial x_i}$.

Полезные материалы :

```
https://kevinzakka.github.io/2016/09/14/batch_normalization/
http://cthorey.github.io./backpropagation/
```