2.1 Change in the learning rate

$$W_t = W_{t-1} - \eta' \cdot \frac{\partial L}{\partial W}$$

$$\eta' = \frac{\eta}{\sqrt{V_t + \varepsilon}}$$ very small values

$$V_t = V_{t-1} + \left(\frac{\partial L}{\partial w}\right)^2$$

$$V_t = \sum_{i=1}^{n} (\partial L / \partial w)^2$$

$$V_t = \sum_{i=1} \frac{\partial L}{\partial w}^2 \qquad \left[\frac{\partial L}{\partial w_1^2} + \frac{\partial L}{\partial w_2^2} \cdots + \frac{\partial L}{\partial w^n}\right]$$

Adaptive adjust :

3.1 Changing in both SC and SR.

R

Date :

(7) ~~root~~ ~~beauty~~ (RMS property).

$$W_t = W_{t-1} - \eta' . \frac{\partial L}{\partial w_{t-1}}$$

~~η' = [illegible]~~

$$\eta' = \eta . \frac{1}{\sqrt{V_t + \epsilon}}$$

$$V_t = \beta . V_{t-1} + (1-\beta) . \left(\frac{\partial L}{\partial w}\right)^2$$

Adam - Adaptive Momentum
RMS prop

Adv. of RMS prop: Updating gradient values simultaneous.

$$W_t = W_{t-1} + \eta' . \frac{\partial L}{\partial w}$$

$$W_t = W_{t-1} + \eta' . m_t$$

$$m_t = \beta . m_{t-1} + (1-\beta) . \frac{\partial L}{\partial w_{t-1}}$$

$$\eta' = \frac{\eta}{\sqrt{v_t + \epsilon}}$$

$$V_t = \beta . V_t + (1-\beta)(\partial L / \partial w)^2$$

$$\hat{m_t} = \frac{m_t}{1 - \beta_1^t} \; ; \; \hat{V_t} = \frac{v_t}{1 - \beta_2^t}$$

2/02/2021

Training tips in Deep Neural Network

Normalization

1) Min-Max Normalization

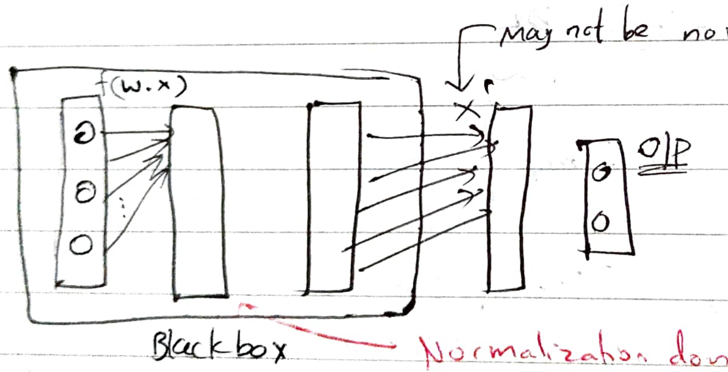
$$\boxed{\frac{x - x_{min}}{x_{max} - x_{min}}}$$ (0-1 range)

2) Z-score Normalization

$N(0,1)$ $\quad \frac{x-\mu}{\sigma}$

$N(\mu, \sigma)$

3) Batch Normalization



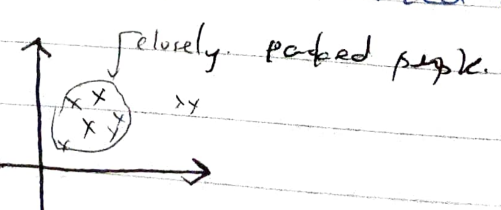
'x' is normalized, but when w.x and non-linear function is used, it (say, x' here) may no longer be normalized.

"Internal covariate shift" occurs when samples of one distribution has difference within the class.

Eg:- Actual class: cats

When plotted: Black and orange cats are separate.

It can be avoided if normalized data is used.



Date :

For mini-batch $B = \{x_1, x_2, \ldots, x_B\}$

o/p: $y = BN_{\gamma, \beta}(x_B)$

$$x^* = BN_{\gamma, \beta}(x_B)$$

x' is considered as I/P to next hidden layer.

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

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

$$\boxed{x_i' = \frac{x_i - \mu_B}{\sqrt{\sigma_B + \varepsilon}}}$$

$$\boxed{x_i' = \underbrace{\gamma}_{\text{Scaling}} x_i' + \underbrace{\beta}_{\text{Shifting}}}$$

OVERFITTING AND UNDERFITTING :-

Mutually exclusive datasets for training, validation and testing.

Solution : regularization

<u>Regularization</u> :-

→ used for better model generalization.

→ General cost function with regularization for training is defined as

Cost function = Loss + Regularization Term.