

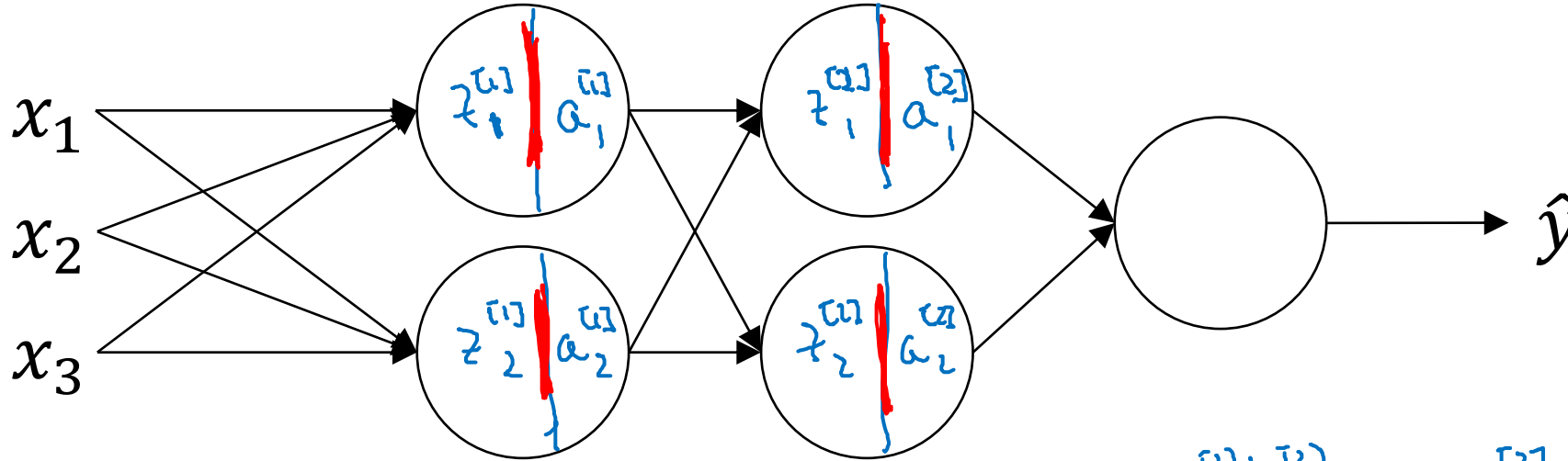


deeplearning.ai

Batch Normalization

Fitting Batch Norm
into a neural network

Adding Batch Norm to a network



$$X \xrightarrow{W^{[1]}, b^{[1]}} \underline{z^{[1]}} \xrightarrow[\text{Batch Norm (BN)}]{\beta^{[1]}, \gamma^{[1]}} \underline{z^{[1]}} \rightarrow a^{[1]} = g(z^{[1]}) \xrightarrow{W^{[2]}, b^{[2]}} \underline{z^{[2]}} \xrightarrow[\text{BN}]{\beta^{[2]}, \gamma^{[2]}} \underline{z^{[2]}} \rightarrow a^{[2]} \rightarrow \dots$$

Parameters: $\left\{ W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}, \dots, W^{[L]}, b^{[L]} \right\}$

$\rightarrow \underline{\beta^{[1]}, \gamma^{[1]}, \beta^{[2]}, \gamma^{[2]}, \dots, \beta^{[L]}, \gamma^{[L]}}$

$\rightarrow \underline{\beta}$

$d\beta^{[L]}$

$\beta = \beta - \alpha d\beta^{[L]}$

tf.nn.batch-normalization ←

Working with mini-batches

$$\underline{X^{[1]}} \xrightarrow{W^{[1]}, b^{[1]}} \underline{z^{[1]}} \xrightarrow[\text{BN}]{\beta^{[1]}, \gamma^{[1]}} \underline{\tilde{z}^{[1]}} \rightarrow g^{[1]}(\tilde{z}^{[1]}) = a^{[1]} \xrightarrow{W^{[2]}, b^{[2]}} \underline{z^{[2]}} \rightarrow \dots$$

$$\boxed{X^{[2]}} \rightarrow \underline{z^{[2]}} \xrightarrow[\text{BN}]{\beta^{[2]}, \gamma^{[2]}} \underline{\tilde{z}^{[2]}} \rightarrow \dots$$

$$X^{[3]} \rightarrow \dots$$

Parameters: $W^{[L]}, \cancel{b^{[L]}}, \beta^{[L]}, \gamma^{[L]}$

\uparrow $(n^{[L]}, 1)$ \uparrow $(n^{[L]}, 1)$ \uparrow $(n^{[L]}, 1)$

\uparrow $z^{[L]}_{(n^{[L]}, 1)}$

$$\rightarrow \underline{z^{[L]}} = W^{[L]} a^{[L-1]} + \cancel{b^{[L]}}$$

\uparrow

$$z^{[L]} = W^{[L]} a^{[L-1]}$$

$$z^{[L]}_{\text{norm}}$$

$$\rightarrow \underline{\tilde{z}^{[L]}} = \gamma^{[L]} z^{[L]}_{\text{norm}} + \boxed{\beta^{[L]}}$$

Implementing gradient descent

for $t = 1 \dots \text{num Mini Batches}$

Compute forward pass on $X^{\{t\}}$.

In each hidden layer, use BN to replace $\underline{z}^{\{t\}}$ with $\underline{\hat{z}}^{\{t\}}$.

Use backprop to compute $\underline{dw}^{\{t\}}$, ~~$\underline{db}^{\{t\}}$~~ , $\underline{dp}^{\{t\}}$, $\underline{df}^{\{t\}}$

Update params $\left. \begin{aligned} w^{\{t\}} &:= w^{\{t-1\}} - \alpha \underline{dw}^{\{t\}} \\ \beta^{\{t\}} &:= \beta^{\{t-1\}} - \alpha \underline{dp}^{\{t\}} \\ \gamma^{\{t\}} &:= \dots \end{aligned} \right\} \leftarrow$

Works w/ momentum, RMSprop, Adam.