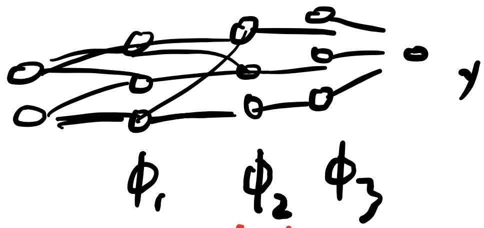
Last time

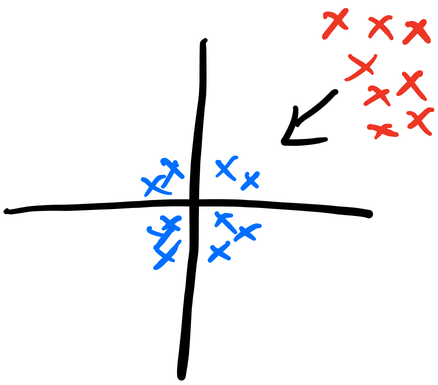
Normalization: keep data at consistent mean and variance

Advantages

$$BN(x) = \frac{x - E[x]}{\sqrt{Var[x]}}$$

$$\left|\frac{dL}{dw_1}\right| = |x|\,|\sigma'(x w_1)| \prod_{i=1}^{L} \left|\frac{d\phi_i}{d\phi_{i-1}}\right|$$

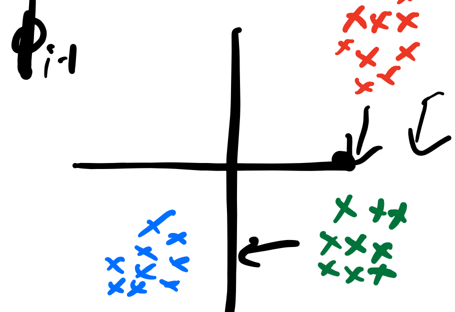




$$\left|\frac{dL}{dw_i}\right| = |\phi_{i-1}|\,|\sigma'(\phi_{i-1} w_i)| \prod_{j=i+1}^{L} \left|\frac{\partial\phi_j}{d\phi_{j-1}}\right|$$

$$BN(\phi_{i-1}) = \frac{\phi_{i-1} - E[\phi_{i-1}]}{\sqrt{Var(\phi_{i-1}) - \varepsilon}}$$

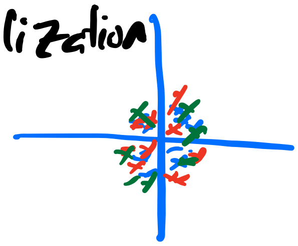
w/o Normalization



$\phi_{i-1} = \sigma(xW)$

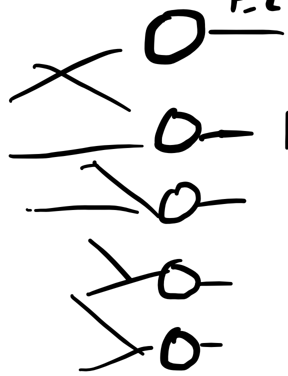
Changes

W/ normalization



Batch norm

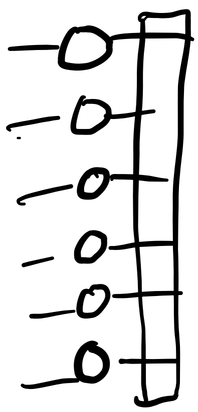
$E[\phi_{:,1}] = 0 \quad Var = 1$



$E[\ ] = 0 \quad Var$

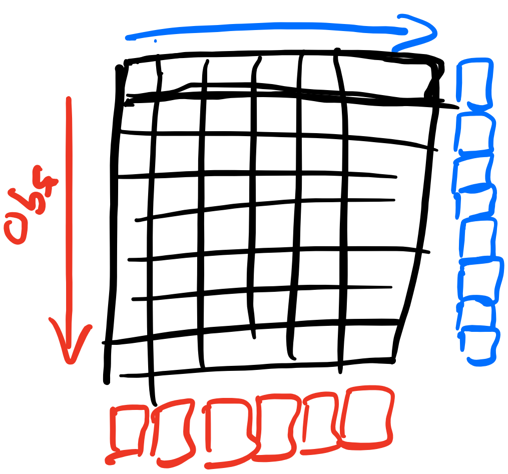
⋮

Layer Norm



$E[x] = 0$

$Var = 1$



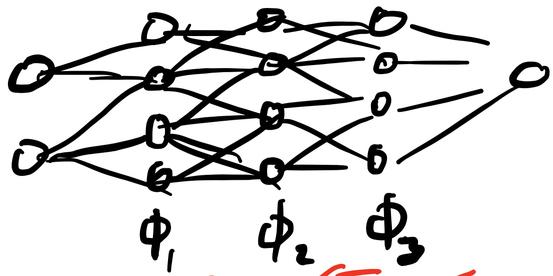
$= X$

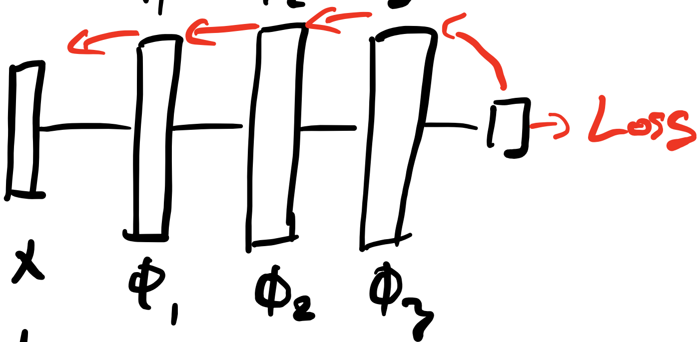
$N \times d$

obj # neurons

Batch norm

Residual Networks

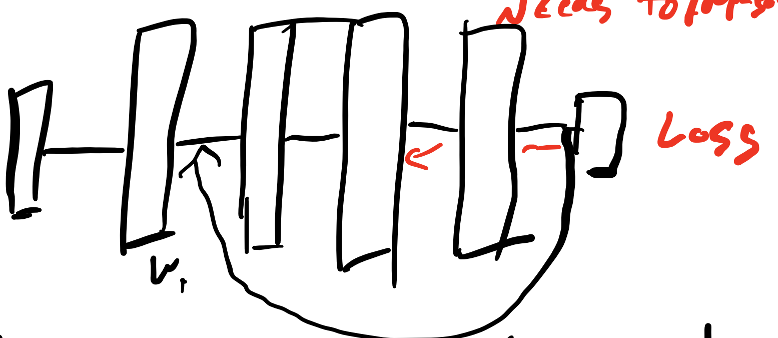




$\phi_1 = \sigma(x w_1)$ $\quad \phi_2 = \sigma(\phi_1 w_2)$ ...

$$\frac{dL}{dw_1} = x\sigma'(x^T w_1)\frac{d\phi_2}{d\phi_1}\cdot\frac{d\phi_3}{d\phi_2}\cdots\frac{dL}{d\phi_L}$$

Needs to propagate Info about Loss



$$\frac{dL}{dw_1} = \underbrace{x\sigma'(xv_1)}\left[\left(\prod_{i=1}^{L}\frac{d\phi_i}{d\phi_{i-1}}\right)\frac{dL}{d\phi_L} + \frac{dL}{d\phi_1}\frac{d\phi_1}{dw_1}\right]$$

Approach: Residual Function

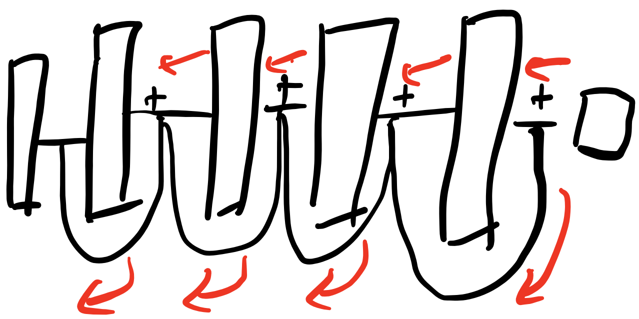
$$\hat{f}(x) \longrightarrow f(x) = \underbrace{\hat{f}(x)} + x$$

Residual Layer

Layer : $\phi_i = \underbrace{\sigma(\phi_{i-1}^T w_i)}_{\text{computes next value}}$

Residual Layer: $\phi_i = \underbrace{\sigma(\phi_{i-1}^T w)}_{\hat{\phi}_i} + \phi_{i-1}$ $\quad \frac{d}{d\phi_{i-1}} = 1$

$\hat{\phi}_i = \sigma(\phi_{i-1} w) = \underbrace{\phi_i - \phi_{i-1}}_{\text{residual}}$



Backprop step

$$\frac{dL}{d\phi_{i-1}} = \frac{dL}{d\phi_i}\cdot\frac{d\phi_i}{d_{i-1}} = \frac{dL}{d\phi_i}\left[\frac{d}{d\phi_{i-1}}\sigma(\phi_{i-1} w_i)\right] + \frac{dL}{d\phi_f}$$

$$= \frac{dL}{d\phi_i}\left[\frac{d}{d\phi_{i-1}}\left(\sigma(\phi_{i-1} w_i) + \phi_{i-1}\right)\right]$$

$$= \frac{dL}{d\phi_i}\left[w_i\sigma'(\phi_{i-1}) + 1\right]$$

w/o residual

$$\frac{dL}{dw_1} = x\sigma'(xw_1)\prod_{i=1}^{L}\left(\frac{d\phi_i}{d\phi_{i-1}}\right)\frac{dL}{d\phi_L}$$

w/ residual

$$\frac{dL}{dw_1} = x\sigma'(xw_1)\prod_{i=1}^{L}\left(\frac{d\hat{\phi}_i}{d\phi_{i-1}}+1\right)\frac{dL}{d\phi_L}$$