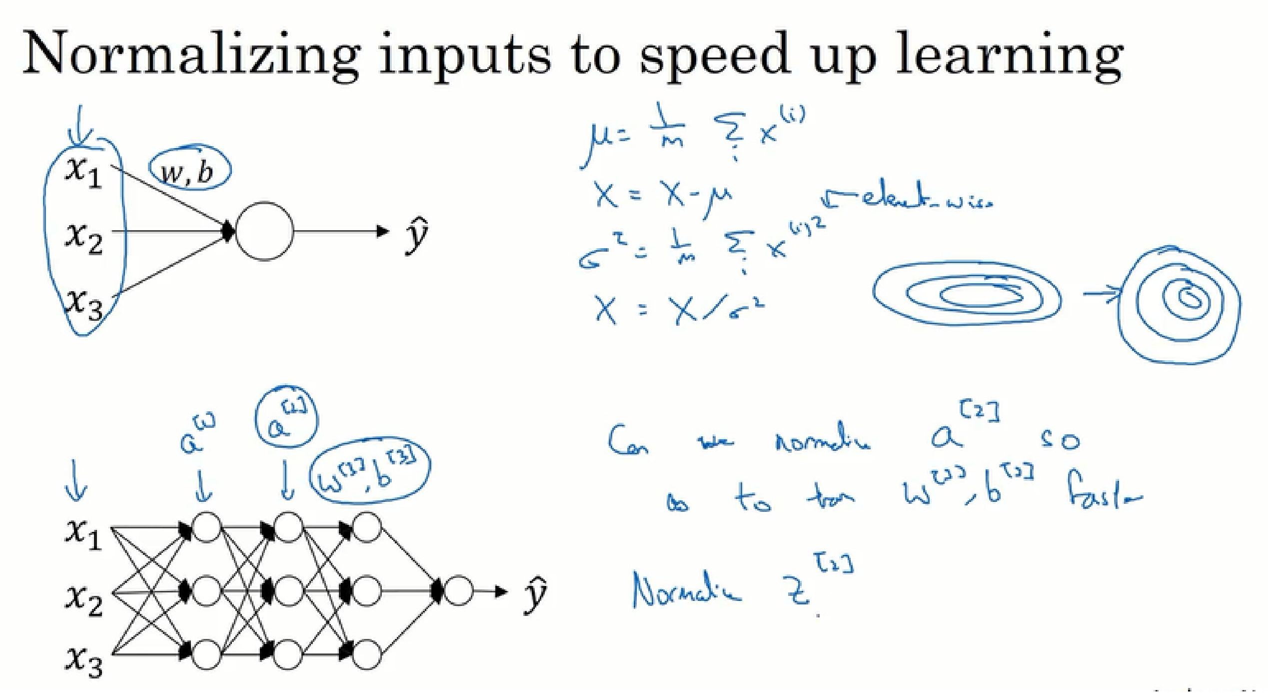
Normalizing activations

- normalize $z^{[l]}$ to train $W^{[l+1]}, b^{[l+1]}$ faster



Implementing batch norm

$$Given\ some\ intermediate\ values\ in\ NN\ \underbrace{z^{(1)}, \cdots, z^{(m)}}_{z^{[l](i)}}$$

$\mu = \frac{1}{m} \sum_i z^{(i)}$

$\sigma^2 = \frac{1}{m} \sum_i (z_i - \mu)^2$

$z_{norm}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$

$\tilde{z}^{(i)} = \gamma z_{norm}^{(i)} + \beta,\ \gamma\ and\ \beta\ are\ learnable\ parameters\ of\ models$

$if: \gamma = \sqrt{\sigma^2 + \epsilon}\ \&\ \beta = \mu,\ then: \tilde{z}^{(i)} = z^{(i)}$

$use\ \tilde{z}^{[l](i)}\ instead\ of z^{[l](i)}$

- batch norm makes the input of hidden units to have **standardized mean and variance**, which are **controlled by learnable parameters γ and β**. These parameters can be **set by the learning algorithm to whatever it wants**.

Applying batch norm to a network

$$X \xrightarrow{W^{[1]}, b^{[1]}} Z^{[1]} \xrightarrow[BatchNorm]{\beta^{[1]}, \gamma[1]} \tilde{Z}^{[1]} \rightarrow a^{[1]} = g^{[1]}(\tilde{z}^{[1]}) \xrightarrow{W^{[2]}, b^{[2]}} Z^{[2]} \tag{1}$$

$parameters:$

$$\begin{cases} W^{[1]}, b^{[1]}, \cdots, W^{[L]}, b^{[L]} \\ \beta^{[1]}, \gamma^{[1]}, \cdots, \beta^{[L]}, \gamma^{[L]} \end{cases} \tag{2}$$

$$\begin{aligned} d\beta^{[l]},\ d\gamma^{[l]} &= backprop \\ \beta^{[l]} &:= \beta^{[l]} - \alpha d\beta^{[l]} \\ \gamma^{[l]} &:= \gamma^{[l]} - \alpha d\gamma^{[l]} \end{aligned} \tag{3}$$

Batch norm working with mini-batches

$$\begin{aligned} X^{\{1\}} \xrightarrow{W^{[1]}, b^{[1]}} Z^{[1]} \xrightarrow[BatchNorm]{\beta^{[1]}, \gamma[1]} \tilde{Z}^{[1]} \rightarrow a^{[1]} = g^{[1]}(\tilde{z}^{[1]}) \xrightarrow{W^{[2]}, b^{[2]}} Z^{[2]} \\ X^{\{2\}} \xrightarrow{W^{[1]}, b^{[1]}} Z^{[1]} \xrightarrow[BatchNorm]{\beta^{[1]}, \gamma[1]} \tilde{Z}^{[1]} \rightarrow a^{[1]} = g^{[1]}(\tilde{z}^{[1]}) \xrightarrow{W^{[2]}, b^{[2]}} Z^{[2]} \\ X^{\{3\}} \rightarrow \cdots \end{aligned} \tag{4}$$

- $In\ mini-batches,\ batch\ norm\ zeros\ out\ the\ bias\ of\ Z$
  - $Z^{[l]} = W^{[l]} A^{[l]},\ b^{[l]} = 0$
  - $Z^{[l]}_{norm}$
  - $\tilde{Z}^{[l]} = A^{[l]} Z^{[l]}_{norm} + \beta^{[l]}$
  - $parameters: W^{[l]}, \beta^{[l]}, \gamma^{[l]}$
  - $Z^{[l]}, \beta^{[l]}, \gamma^{[l]}\ are\ (n^{[l]}, 1)$

Implement batch norm

$for\ t = 1\ \cdots\ num\ of\ mini-batches$

$compute\ forwardprop\ on\ X^{\{t\}}$

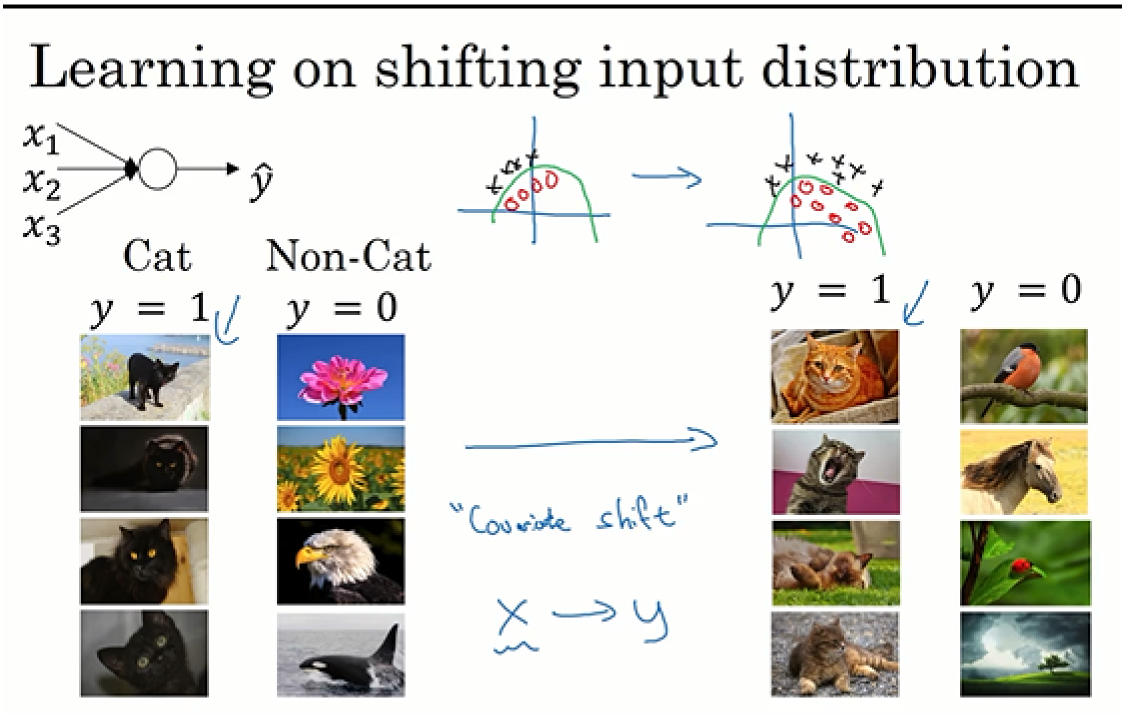
$In\ each\ hidden\ layer,\ use\ BatchNorm\ to\ get\ \tilde{Z}^{[l]}\ from\ Z^{[l]}$

$Use\ backprop\ to\ compute\ dW^{[l]}, d\beta^{[l]}, d\gamma^{[l]}$

$Update\ parameters$

Covariate shift

- the distribution of training X and Y change, even though the function do the same work



- Batch norm makes the hidden units` output have mean and variance governed by β and γ .

Batch norm at test time

$\mu, \sigma^2 :\ estimate\ using\ exponentially\ weighted\ average\ across\ mini-batch$

$X^{\{1\}} \rightarrow \mu^{\{1\}[l]}, (\sigma^{\{1\}[l]})^2$

$X^{\{2\}} \rightarrow \mu^{\{2\}[l]}, (\sigma^{\{2\}[l]})^2$

$\cdots\ \cdots$

$Z_{norm} = \frac{Z-\mu}{\sqrt{\sigma^2+\epsilon}}$

$\tilde{Z} = \gamma Z_{norm} + \beta$