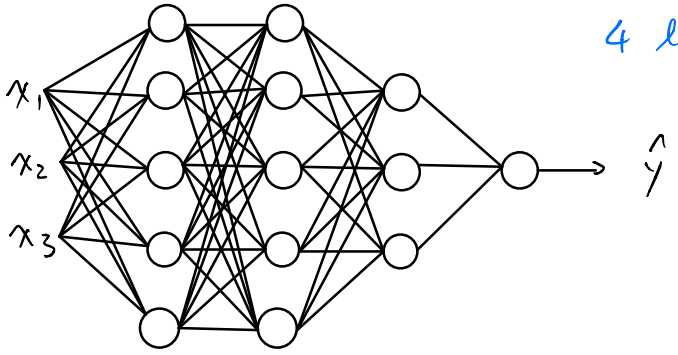


4 layer NN

$L = 4$ (# layers)

$n^{[l]}$ = # units in layer l

$a^{[l]}$ = activations in layer l

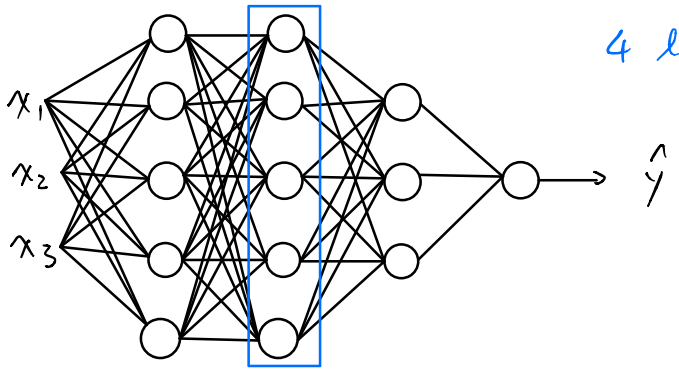
$a^{[l]} = g^{[l]}(z^{[l]})$ $\quad (n^{[l]}, m)$ — m for all samples

$w^{[l]}$ = weights for $z^{[l]}$: $(n^{[l]}, n^{[l-1]})$

$$\begin{cases} z^{[l]} = w^{[l]} a^{[l-1]} + b^{[l]} \\ a^{[l]} = g^{[l]}(z^{[l]}) \end{cases}$$

$\Rightarrow$ Vectorized:

$$\begin{cases} Z^{[l]} = W^{[l]} A^{[l-1]} + b^{[l]} \\ A^{[l]} = g^{[l]}(Z^{[l]}) \end{cases}$$



4 layer NN

layer l: $w^{[l]}$, $b^{[l]}$

Forward: Input $a^{[l-1]}$, output $a^{[l]}$

$z^{[l]} = w^{[l]} a^{[l-1]} + b^{[l]}$ ⇐ cache $z^{[l]}$

$a^{[l]} = g^{[l]}(z^{[l]})$

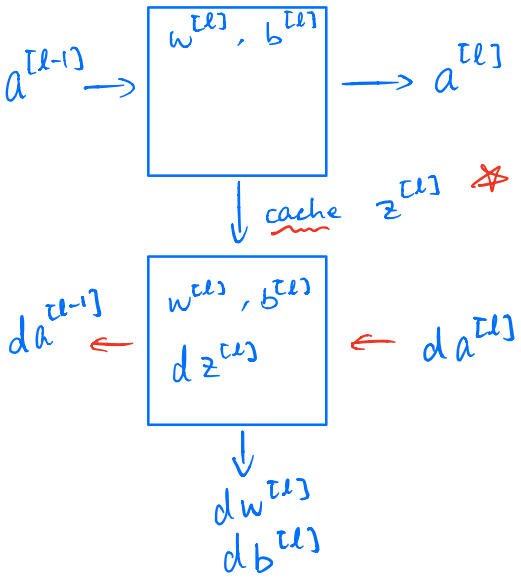
Backward: Input $da^{[l]}$, output $da^{[l-1]}$

cache ($z^{[l]}$)

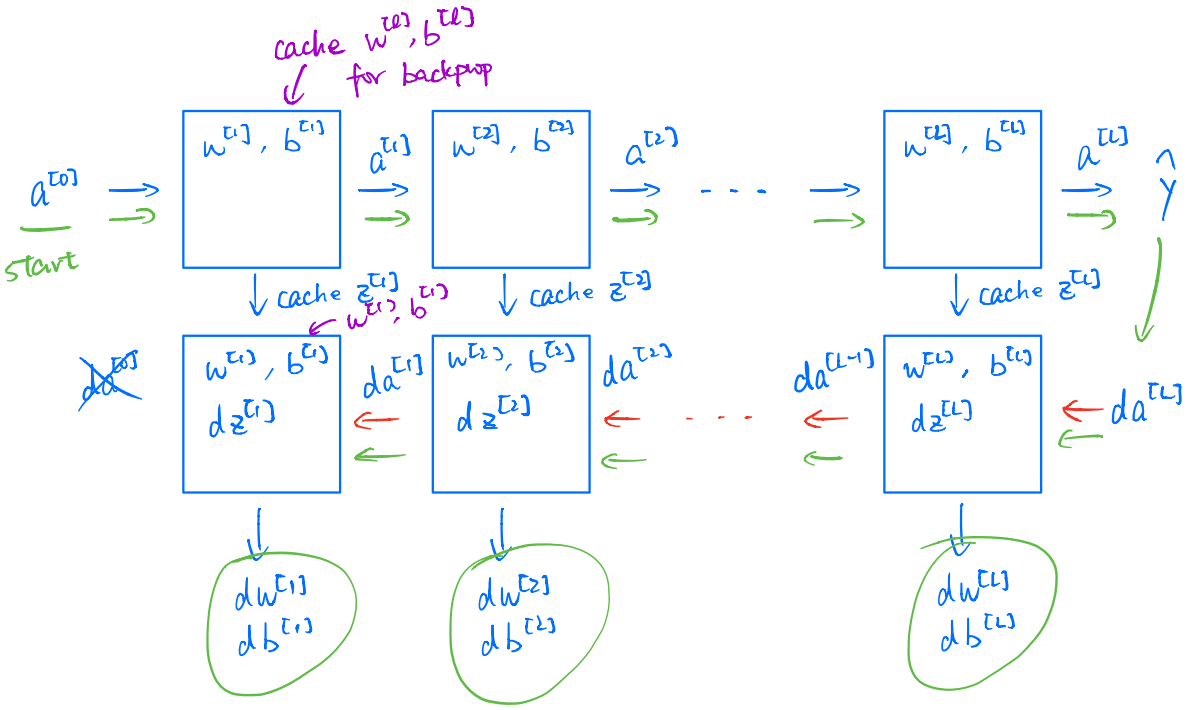
$dw^{[l]}$

$db^{[l]}$

layer l



Forward and Backward Functions



$$w^{[l]} = w^{[l]} - \alpha\, dw^{[l]}$$

$$b^{[l]} = b^{[l]} - \alpha\, db^{[l]}$$

Forward

Input $a^{[l-1]}$ (also $w^{[l]}, b^{[l]}$)

Output $a^{[l]}$, cache $z^{[l]}$

$$\Rightarrow \begin{cases} z^{[l]} = w^{[l]} a^{[l-1]} + b^{[l]} \\ a^{[l]} = g^{[l]}(z^{[l]}) \end{cases}$$

Practice !!

Backward

Input $da^{[l]}$

Output $da^{[l-1]}$, $dw^{[l]}$, $db^{[l]}$

$$\Rightarrow \begin{cases} dz^{[l]} = da^{[l]} * g'^{[l]}(z^{[l]}) \\ dw^{[l]} = dz^{[l]} \cdot a^{[l-1]T} \\ db^{[l]} = dz^{[l]} \\ da^{[l-1]} = w^{[l]T} dz^{[l]} \end{cases}$$

Vectorized

Forward

Input $A^{[l-1]}$ (also $W^{[l]}, b^{[l]}$)

Output $A^{[l]}$, cache $Z^{[l]}$

$$\Rightarrow \begin{cases} Z^{[l]} = W^{[l]} A^{[l-1]} + b^{[l]} \\ A^{[l]} = g^{[l]}(Z^{[l]}) \end{cases}$$

Backward

Input $dA^{[l]}$

Output $dA^{[l-1]}$, $dW^{[l]}$, $db^{[l]}$

$$\Rightarrow \begin{cases} dZ^{[l]} = dA^{[l]} * g^{[l]\prime}(Z^{[l]}) \\ dW^{[l]} = \frac{1}{m} dZ^{[l]} \cdot A^{[l-1]T} \\ db^{[l]} = \frac{1}{m} np.sum(dZ^{[l]}, axis=1, keepDim = True) \\ dA^{[l-1]} = W^{[l]T} dZ^{[l]} \end{cases}$$

Practice

$$dZ^{[l]} = dA^{[l]} * \frac{\partial A^{[l]}}{\partial Z^{[l]}} \quad \left(\frac{\partial A^{[l]}}{\partial Z^{[l]}} = g^{[l]\prime}(Z^{[l]})\right)$$

$(n^{[l]}, m) \quad (n^{[l]}, m) \quad (n^{[l]}, m)$

$$dW^{[l]} = \frac{1}{m} dZ^{[l]} \cdot A^{[l-1]T}$$

$(n^{[l]}, n^{[l-1]}) \quad (n^{[l]}, m) \quad (m, n^{[l-1]})$

$$db^{[l]} = \frac{1}{m} dZ^{[l]}$$

$(n^{[l]}, 1)$ ← sum axis = 1

$$dA^{[l-1]} = \frac{\partial Z^{[l]}}{\partial A^{[l-1]}} \cdot dZ^{[l]} \quad \left(\frac{\partial Z^{[l]}}{\partial A^{[l-1]}} = W^{[l]T}\right)$$

$(n^{[l-1]}, 1) \quad (n^{[l]}, n^{[l-1]}) \quad (n^{[l]}, 1)$

Hyper parameters

Parameters: $W^{[1]}$, $b^{[1]}$, $W^{[2]}$, $b^{[2]}$, ...

Hyperparameters: learning rate α

iterations

hidden layers L

hidden units $n^{[1]}$, $n^{[2]}$, ...

Choice of activation function

Later: Momentum

mini batch size

regularizations

...

Try different values → empirical process