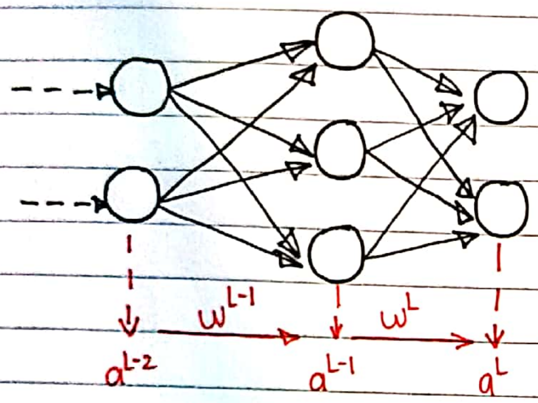
Backpropagation in neural network explained:

↳ Let say we have the following neural network:



→ Let say we have the activated outputs at each layer are a^L, a^{L-1} and a^{L-2}

* Activated output = output after activation.

→ Example: To calculate activated output at layer L, we have:

$$(*)\begin{cases} z^L = a^{L-1} \cdot w^L & (1) \\ a^L = \sigma(z^L) & (2) \end{cases}$$

·) Where z^L in (1) calculates unactivated output, a^{L-1} is activated output in previous layer, w^L is the set of weights at layer L. In (2), $\sigma(\cdot)$ represents the activation function.

(1) Gradients:

↳ To perform gradient descent, we need the gradient of a loss function L with respect to all sets of weights $w^L, w^{L-1}, \ldots$

↳ Let say our loss function is a simple squared difference loss, let's calculate the gradients for w^L.

·) $\dfrac{\partial L}{\partial w^L} = \dfrac{\partial}{\partial w^L}\left[(y-a^L)^2\right]$ (where y is the labels).

·) $\dfrac{\partial L}{\partial w^L} = \dfrac{\partial L}{\partial a^L} \cdot \dfrac{\partial a^L}{\partial z^L} \cdot \dfrac{\partial z^L}{\partial w^L}$ (chain rule in differentiation).

↳ Let's examine each component:

(1) $\dfrac{\partial L}{\partial a^L} = \dfrac{\partial}{\partial a^L}\left[(y-a^L)^2\right] = -2(y-a^L) = 2(a^L-y)$.

(2) $\dfrac{\partial a^L}{\partial z^L} = \dfrac{\partial}{\partial z^L}\left[\sigma(z^L)\right] = \sigma'(z^L)$.

(3) $\dfrac{\partial z^L}{\partial w^L} = \dfrac{\partial}{\partial w^L}\left[(a^{L-1} \cdot w^L)\right] = a^{L-1}$

⇒ From the above: $\dfrac{\partial L}{\partial w^L} = 2(a^L-y)\,\sigma'(z^L)\,a^{L-1}$ #

② Back-propagation.

↳ Okay now we know $\frac{\partial L}{\partial w^L}$. However, we need to know $\frac{\partial L}{\partial w^k}$ with k being any layer in our neural network. Let's state the formula again:

$$\Rightarrow \frac{\partial L}{\partial w^k} = \frac{\partial L}{\partial a^k} \cdot \frac{\partial a^k}{\partial z^k} \cdot \frac{\partial z^k}{\partial w^k} \quad (4)$$

↳ We know $\frac{\partial a^k}{\partial z^k} = \sigma'(z^k)$ is just the derivative of the activation function and $\frac{\partial z^k}{\partial w^k} = a^{k-1}$ will just always be the output from previous layer.

↳ So to calculate (4) we need $\frac{\partial L}{\partial a^k}$, let's find it using induction:

1. Base case: $\frac{\partial L}{\partial a^L} = 2(a^L - y)$

* We already know this

2. $$\frac{\partial L}{\partial a^{L-1}} = \frac{\partial L}{\partial a^L} \cdot \frac{\partial a^L}{\partial z^L} \cdot \frac{\partial z^L}{\partial a^{L-1}} = \frac{\partial L}{\partial a^L} \cdot \sigma'(z^L) \cdot w^L$$

* We know this

3. $$\frac{\partial L}{\partial a^{L-2}} = \frac{\partial L}{\partial a^{L-1}} \cdot \frac{\partial a^{L-1}}{\partial z^{L-1}} \cdot \frac{\partial z^{L-1}}{\partial a^{L-2}} = \frac{\partial L}{\partial a^{L-1}} \cdot \sigma'(z^{L-1}) \cdot w^{L-1}$$

$$\Rightarrow \boxed{\frac{\partial L}{\partial a^k} = \frac{\partial L}{\partial a^{k+1}} \cdot \sigma'(a^{k+1}) \cdot w^{k+1} \quad (\text{for } k < L)}$$

$\Rightarrow$ Final note: Since as long as we know $\frac{\partial L}{\partial a^{k+1}}$, we can always iteratively propagate our gradients backwards to calculate the gradients for previous layers, it is called "Back-propagation".