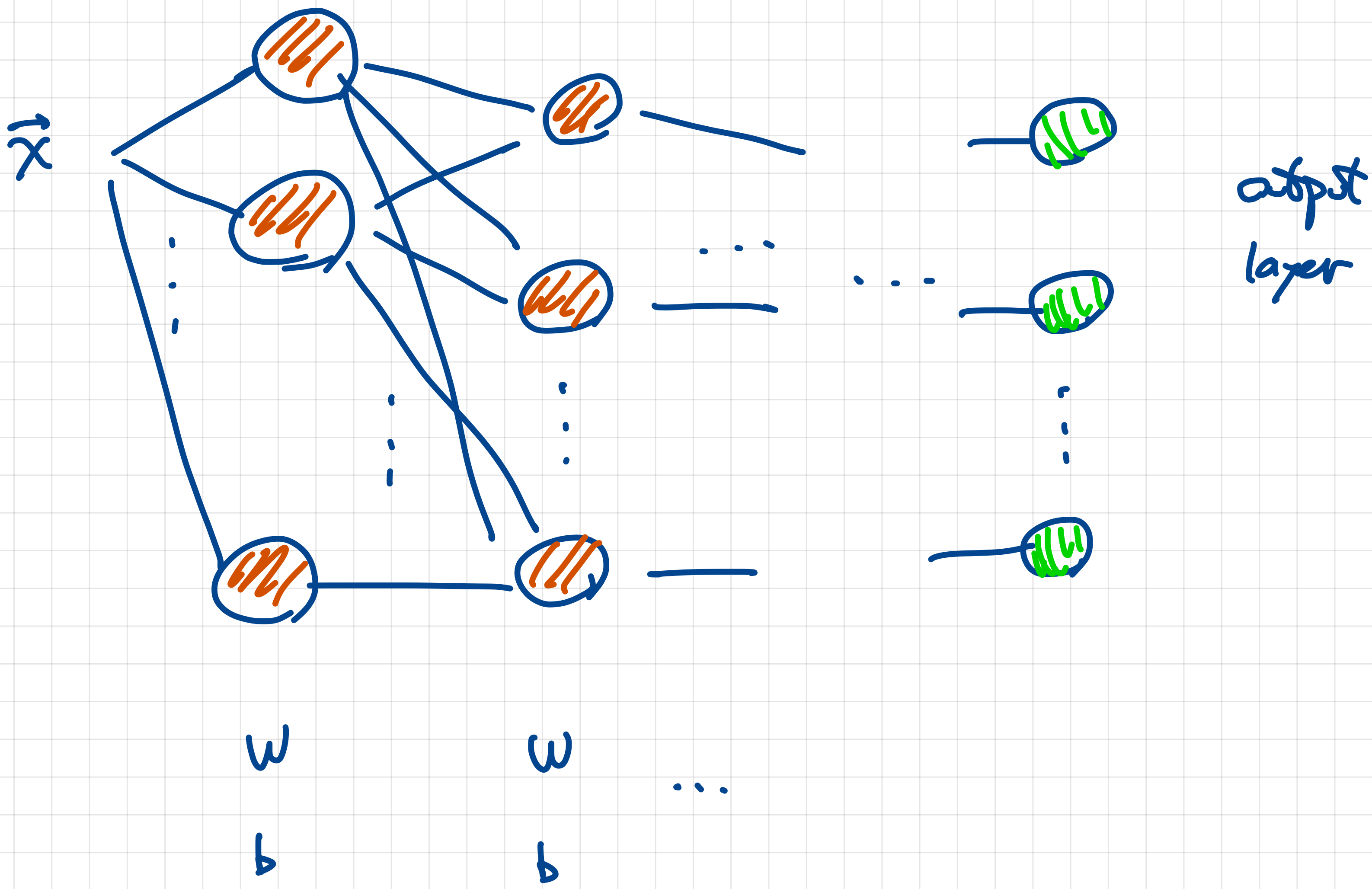
Short Takes 331

Machine learning: Backpropagation

Machine learning: backpropagation.

- A neural network ...



... is ultimately a function designed in a particular way (layers, activation functions, etc) with parameters (weights & biases) to be "learned".

- One defines a cost function to be minimized via stochastic gradient descent.

$$C(w,b) = \sum_{i=1}^{N} C_i(w,b),$$

given data $\{(\vec{x},\vec{y})_1, (\vec{x},\vec{y})_2, \dots, (\vec{x},\vec{y})_N\}$

e.g. $C_i = (\vec{y}_i - \vec{a}(\vec{x}_i, w, b))^2$

$\to$ We need $\vec{\nabla} C_i$!

Let's see...

$$\frac{\partial C}{\partial w_{ij}^{(k)}} = \sum_m \frac{\partial C}{\partial a_m} \frac{\partial a_m}{\partial w_{ij}^{(k)}} \quad , \qquad \text{but...}$$

layer

neuron index in layer

input index

network output

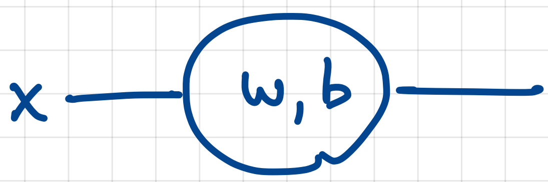
$$\vec{a} = f\Big(\dots \underbrace{f\big(w^{(2)} f(\underbrace{w^{(1)}\vec{x} + \vec{b}^{(1)}}_{z^{(1)}}) + \vec{b}^{(2)}\big)}_{z^{(2)}} + \vec{b}^{(3)} \dots\Big)$$

- We won't tackle this fully, but work out two simple cases...

$$C = C(a)$$

$$a = a(w)$$

$$= f(\underbrace{wx + b}_{z})$$



$a = f(z)$

$z = wx + b$

$$\Rightarrow \quad \frac{\partial C}{\partial w} = \frac{\partial C}{\partial a} \frac{\partial a}{\partial w}$$

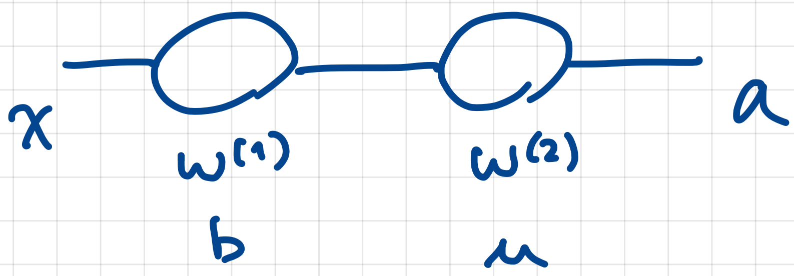
$$\frac{\partial a}{\partial w} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial w}$$

$$\frac{\partial z}{\partial w} = x$$

Note: $\frac{\partial f}{\partial z}$ is fully determined by the activation function and $z = wx + b$.

- Slightly more complicated: <u>two layers</u>

$$a(w) = \underbrace{f\big(w^{(2)} f(\overbrace{w^{(1)}x + b}^{z^{(1)}}) + u\big)}_{z^{(2)}}$$



$\Rightarrow$ • $$\frac{\partial C}{\partial w^{(2)}} = \frac{\partial C}{\partial a}\frac{\partial a}{\partial w^{(2)}}$$

$$\frac{\partial a}{\partial w^{(2)}} = \frac{\partial f}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial w^{(2)}}$$

fully determined by layer 2

$$\frac{\partial z^{(2)}}{\partial w^{(2)}} = f(z^{(1)}) \quad \checkmark$$

$$= \frac{\partial C}{\partial a}\frac{\partial f}{\partial z^{(2)}} f(z^{(1)})$$

!

- $$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a}\frac{\partial a}{\partial w^{(1)}}$$

$$\frac{\partial a}{\partial w^{(1)}} = \frac{\partial f}{\partial z^{(2)}}\frac{\partial z^{(2)}}{\partial w^{(1)}}$$

$$\frac{\partial z^{(2)}}{\partial w^{(1)}} = w^{(2)}\frac{\partial f}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial w^{(1)}}$$

$$\frac{\partial z^{(1)}}{\partial w^{(1)}} = x$$

$$= \frac{\partial C}{\partial a}\frac{\partial f}{\partial z^{(2)}} w^{(2)}\frac{\partial f}{\partial z^{(1)}} \cdot x$$

!

The gist of back propagation is that we perform far fewer operations if we calculate the gradient of the last layer first and then use the intermediate steps to calculate the gradient of the previous layer, and so on.

If we did forward propagation, we would be computing the same elements many times.