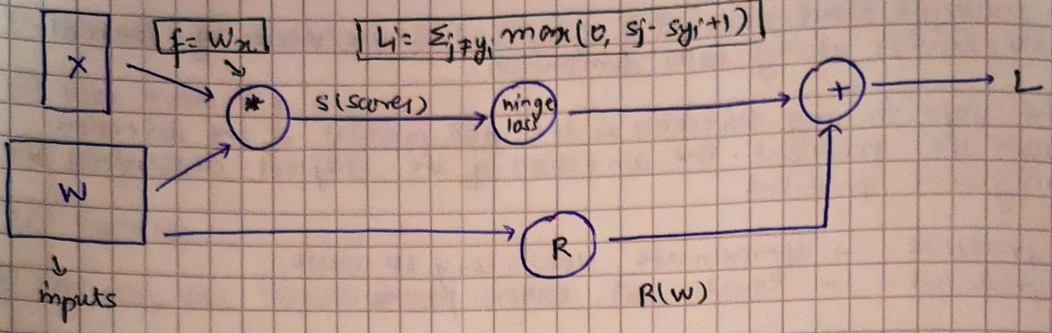
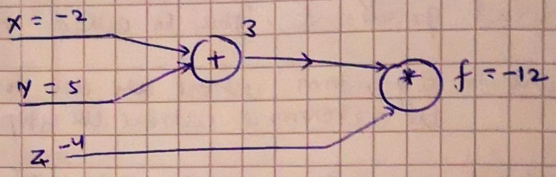
Backpropogation and Neural Networks

① Calculating analytic gradient with computation graph.



② Backpropogation.

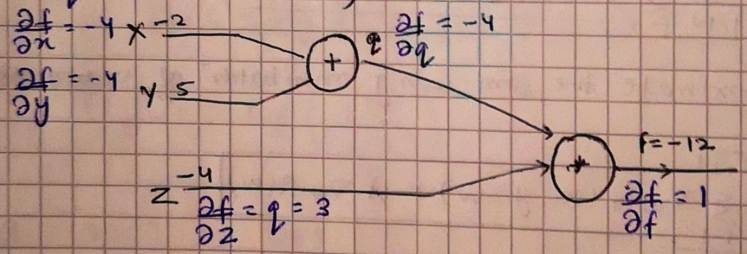
$f(x,y,z) = (x+y)z$



$q = x + y$ $\quad \frac{\partial q}{\partial x} = 1$ $\quad \frac{\partial q}{\partial y} = 1$

$f = qz$ $\quad \frac{\partial f}{\partial q} = z$ $\quad \frac{\partial f}{\partial z} = q$

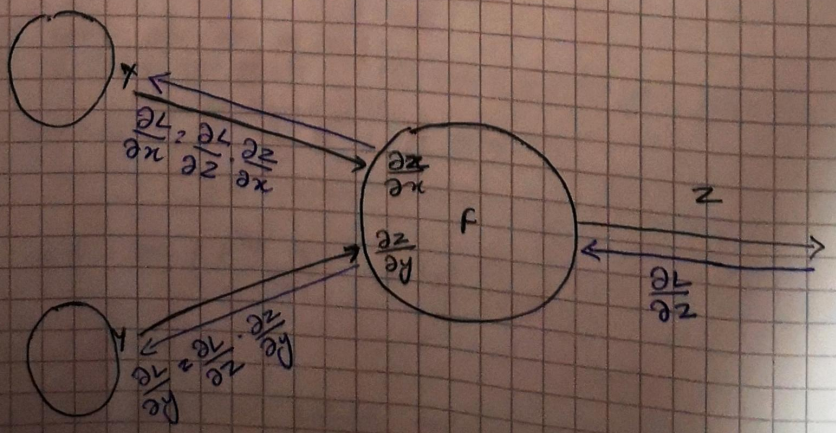
} we want $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ and $\frac{\partial f}{\partial z}$



using chain rule for calculating $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$

$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial y} = -4 \times 1 = \underline{-4}$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x} = -4 \times 1 = \underline{-4}$



① Patterns in Backward flow.

→ Just as an add gate acts as a "gradient distributer" i.e. distributes the upstream gradient equally

→ A max gate acts as a "gradient router" meaning one of the node will get the full value of upstream gradient and the other will not receive anything. We can use this to route our desired output

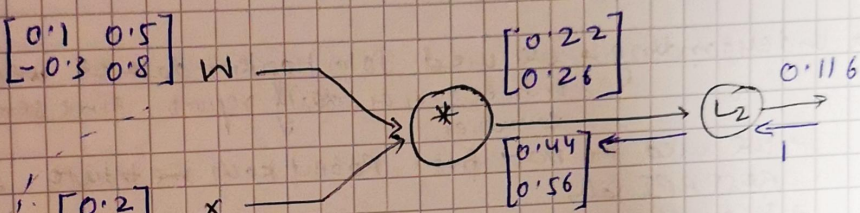
→ A multiplier gate acts as a gradient switcher.

② Vectorized example.

$$f(x, W) = \|W \cdot x\|^2 = \sum_{i=1}^{n} (W \cdot x)_i^2$$

$\in R^n$ $\in R^{n \times n}$

$$\begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix} W \qquad \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} x$$



$$\begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix} \qquad \begin{bmatrix} 0.44 \\ 0.56 \end{bmatrix}$$

$$q = W \cdot x = \begin{pmatrix} w_{1,1} x_1 + \cdots + w_{1,n} x_n \\ w_{n,1} x_1 + \cdots + w_{n,n} x_n \end{pmatrix}$$

$$f(q) = \|q\|^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i$$

$$\nabla_q f = 2q$$

$$\frac{\partial q_k}{\partial W_{i,j}} = 1_{k=i} \, x_j$$

$$\frac{\partial f}{\partial W_{i,j}} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}}$$

$$= \sum (2q_k)(1_{k=i} \, x_j)$$

$$= 2q_i x_j$$

$$\nabla_W f = 2q \cdot x^T$$

Eg. $\begin{bmatrix} 0.44 \\ 0.52 \end{bmatrix} \begin{bmatrix} 0.2 & 0.4 \end{bmatrix}$

$$= \begin{bmatrix} 0.088 & 0.104 \\ 0.176 & 0.208 \end{bmatrix} = W$$

$$\frac{\partial q_k}{\partial x_i} = W_{k,i}$$

$$\frac{\partial f}{\partial x_i} = \sum \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i}$$

$$= \sum_k 2q_k W_{k,i}$$

$$\nabla_x f = 2W^T \cdot q$$

Eg. ~~[illegible]~~

Eg. $2 \begin{bmatrix} 0.1 & -0.3 \\ 0.5 & 0.8 \end{bmatrix} \begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$

$$= \begin{bmatrix} 0.2 & -0.6 \\ 1.0 & 1.6 \end{bmatrix} \begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$$

$$= \begin{bmatrix} -0.112 \\ 0.636 \end{bmatrix} = X$$