$$X \in R^{2\times 3} \qquad W \in R^{3\times 5}$$

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix} \cdot \begin{pmatrix} W_{11} & W_{12} & W_{13} & W_{14} & W_{15} \\ W_{21} & W_{22} & W_{23} & W_{24} & W_{25} \\ W_{3} & W_{32} & W_{33} & W_{34} & W_{35} \end{pmatrix}$$

$$Y \in R^{2\times 5}$$

$$= \begin{pmatrix} y_{11} & y_{12} & y_{13} & y_{14} & y_{15} \\ y_{21} & y_{22} & y_{23} & y_{24} & y_{25} \end{pmatrix}$$

$$\begin{cases} y_{12} = x_{11}W_{12} + x_{12}W_{22} + x_{13}W_{32} \\ y_{22} = x_{21}W_{12} + x_{22}W_{22} + x_{23}W_{32} \end{cases}$$

$$\begin{cases} \dfrac{\partial y_{12}}{\partial W_{12}} = x_{11} \\[2ex] \dfrac{\partial y_{22}}{\partial W_{22}} = x_{21} \end{cases}$$

J_W has the same shape as W $(\mathbb{R}^{3\times5})$

Assuming $L = f(y) \in \mathbb{R}$

$$J_{W_{12}} = \frac{\partial L}{\partial W_{12}} = \frac{\partial L}{\partial y_{12}} \cdot \underbrace{\frac{\partial y_{12}}{\partial W_{12}}}_{X_{11}} + \frac{\partial L}{\partial y_{22}} \cdot \underbrace{\frac{\partial y_{22}}{\partial W_{12}}}_{X_{21}}$$

$$\begin{cases} \dfrac{\partial y_{12}}{\partial W_{12}} = X_{11} \\ \dfrac{\partial y_{22}}{\partial W_{22}} = X_{21} \end{cases}$$

Assuming $J_y = \frac{\partial L}{\partial y} \in \mathbb{R}^{2\times5}$, the same shape as y

then $J_{W_{12}} = J_{y_{12}} \cdot X_{11} + J_{y_{22}} \cdot X_{21}$

a dot product between 2^{nd} column of J_y and 1^{st} column of X

$$\underset{2\times5}{y} = \underset{2\times3}{X} \; \underset{3\times5}{W}$$

$$\underset{3\times5}{J_W} = \underset{3\times2}{X^T} \cdot \underset{2\times5}{J_y}$$

similarly :

$$\underset{2\times3}{J_X} = \underset{2\times5}{J_y} \cdot \underset{5\times3}{W^T}$$

In code, we refer

J_y : y_grad

J_W : w_grad

$y = a \cdot b$

$$\frac{\partial y}{\partial a} = b$$

$L = f(y)$

$y = a + b$

$$\frac{\partial L}{\partial a}$$

$$\underset{2\times5}{y} = \underset{2\times5}{X} + \underset{2\times5}{b}$$

$J_X = J_y$

$J_b = J_y$

if b is 1×5.

then $\underset{1\times5}{J_b}$ is $\sum_{i=1}^{2} J_{y_i}$ ← row of y

$y = a + b$

$$\frac{\partial y}{\partial a} = 1 \implies \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial a} = \frac{\partial L}{\partial y}.$$

For element-wise computation like ReLU

$$\mathrm{ReLU}(x) = \max(0, x)$$

the gradient is element-wise mapping

Check out matrix cookbook for gradient of other linear algebra operation

Tips :

the gradient of mapping $f: R^m \to R^n$

has the shape of R^m