JACOBIAN (OPTIONAL MATERIAL)

JACOBIAN IS THE DERIVATIVE OF MULTI INPUT - MULTI OUTPUT FUNCTIONS

$$\bar{y} = f(\bar{x})$$

DEFINITION (FOR VECTOR IN - VECTOR OUT)

$$\frac{\partial f(\bar{x})}{\partial \bar{x}} = \left.\begin{array}{c} \\ \\ \\ \end{array}\right\}\!\underset{\text{NO OF OUTPUTS}}{} \quad \underbrace{\begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1}, & \frac{\partial f_1(x)}{\partial x_2}, & \cdots, & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & & \\ \frac{\partial f_m(x)}{\partial x_1}, & \frac{\partial f_m(x)}{\partial x_2}, & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}}_{\text{NO. OF INPUTS}}$$

SOMETIMES THE JACOBIAN IS DEFINED AS A TRANSPOSE OF THIS. IT DOESN'T MATTER UNTIL IT IS CONSISTENT.

FOR MATRIX - VECTOR MULTIPLICATION

$$\bar{y} = \underline{A}\,\bar{x} \quad \rightarrow \quad \bar{y} = f(\bar{x})$$

NO OF INPUTS: NO. OF ELEMENTS OF X

NO OF OUTPUTS: NO OF ELEMENTS OF Y

$$\frac{d\underline{A}\bar{x}}{d\bar{x}} = \underbrace{\begin{bmatrix} \quad \end{bmatrix}}_{\text{NO OF x}} \Big\}\ \text{NO OF Y}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\underline{A}\,\bar{x} = \underbrace{\left\{ \overbrace{\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}}^{\text{NO OF X}} \right.}_{\text{NO OF Y}} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$y_i = a_{i1} x_1 + a_{i2} x_2 + \dots a_{in} x_n$$

$$\frac{dy_i}{dx_j} = a_{ij}$$

$$\Downarrow$$

$$\frac{d\underline{A}\bar{x}}{d\bar{x}} = \underline{A}$$

JACOBIAN WITH RESPECT TO THE MATRIX IS A 3D TENSOR. FOR A MATRIX INPUT – MATRIX OUTPUT FUNCTIONS, IT IS A 4D TENSOR.

$$\underset{[m]}{\bar{y}} = \underset{[m \times n]}{\underline{A}} \; \underset{[n]}{\bar{x}}$$

$$\frac{\partial \bar{y}}{\partial \underline{A}} = \text{TENSOR OF } \underbrace{m}_{\text{NO OF OUTPUTS}} \times \underbrace{m \times n}_{\text{NO OF INPUTS}}$$

$$\frac{\partial \bar{y}}{\partial \underline{A}}_{ijk} = \frac{\partial y_i}{\partial a_{jk}}$$

BUT FORTUNATELY WE ALWAYS HAVE A SCALAR OUTPUT FUNCTION, BECAUSE THE LOSS IS ALWAYS A SCALAR.

NOTE THAT WE ARE ALWAYS CALCULATING THE GRADIENT WITH RESPECT TO THE LOSS, WHICH IS SCALAR.

FOR BATCHED TRAINING, DATA IS A $[b \times n]$ MATRIX. EACH TRANSFORMATION CAN CHANGE THE NUMBER OF CHANNELS, BUT NOT THE NUMBER OF BATCHES, THUS THE OUTPUT IS $[b \times m]$.

FINALLY, WE HAVE THE REST OF THE NETWORK, THAT REDUCE ALL DATA TO A SINGLE SCALAR.

$$\underbrace{\underset{[b \times n]}{x} \longrightarrow \boxed{f}}_{y=f(x)} \longrightarrow \underbrace{\underset{[b \times m]}{y} \longrightarrow \dots \longrightarrow \underset{[1]}{E}}_{E=g(y)}$$

$y = f(x) \qquad E = g(y)$ – THE REST OF THE NETWORK, INCLUDING LOSS.

$$\frac{\partial f(x)}{\partial x} : \underbrace{\overset{i}{b} \times \overset{j}{m}}_{\text{OUTPUT}} \times \underbrace{\overset{k}{b} \times \overset{l}{n}}_{\text{INPUT}}$$

$$\frac{\partial E}{\partial y} : \underbrace{1}_{\text{OUTPUT}} \times \underbrace{\overset{i}{b} \times \overset{j}{m}}_{\text{INPUT}} \sim b \times m$$

WE CAN IGNORE THE 1, IT IS ALWAYS 1 FOR SCALAR OUTPUT

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial y} \underbrace{\otimes_2}_{\text{GENERALIZED MATRIX MULTIPLICATION}} \frac{\partial y}{\partial x}$$

HOW MANY DIMENSIONS TO REDUCE (SUM OVER)

$$[i, j] \otimes_2 [i, j, k, l] = [k, l]$$

IN OUR CASE

$$\underbrace{\underline{X}}_{[b \times n]} \; \underbrace{\underline{A}}_{[n \times m]} = \underbrace{\underline{y}}_{[b \times m]} \qquad E = g(\underline{y})$$

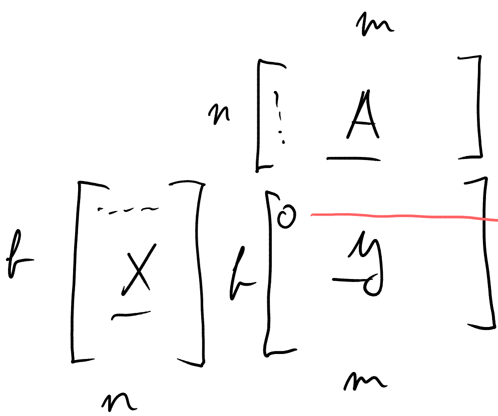
WE ARE INTERESTED IN $\frac{\partial E}{\partial A}$ AND $\frac{\partial E}{\partial X}$.

$\frac{\partial E}{\partial y}$: $\underbrace{1}_{\text{OUT}} \times \underbrace{b \times m}_{\text{IN}} \sim b \times m$ MATRIX

$\frac{\partial y}{\partial A}$: $\underbrace{b \times m}_{\text{OUT}} \times \underbrace{n \times m}_{\text{IN}}$ TENSOR

$\frac{\partial E}{\partial A}$: $\underbrace{1}_{\text{OUT}} \times \underbrace{n \times m}_{\text{IN}} \sim n \times m$ ← WE ARE INTERESTED IN THIS

CHAIN RULE : $\frac{\partial E}{\partial A} = \frac{\partial E}{\partial y} \otimes_2 \frac{\partial y}{\partial A}$



$$y_{ij} = \sum_k X_{ik} A_{kj}$$

↑ DEF OF MATMUL

$$\left(\frac{\partial y}{\partial A}\right)_{ijkl} = \frac{\partial y_{ij}}{\partial A_{kl}} = \begin{cases} 0 & j \neq l \\ X_{ik} & j = l \end{cases}$$

$$\frac{\partial y}{\partial A}_{ijkl} = \begin{cases} 0 & j \neq l \\ X_{ik} & j = l \end{cases}$$

$$\frac{\partial E}{\partial A} = \underset{[k\,l]}{\frac{\partial E}{\partial y}} \otimes_2 \underset{[k\,l\,i\,j]}{\frac{\partial y}{\partial A}}$$

$$\frac{\partial E}{\partial A}_{ij} = \sum_k \underbrace{\sum_l \frac{\partial E}{\partial y}_{kl} \frac{\partial y}{\partial A}_{klij}}_{0 \text{ IF } l \neq j \Rightarrow l = j}$$

$$\frac{\partial E}{\partial A}_{ij} = \sum_k \frac{\partial E}{\partial y}_{kj} \frac{\partial y}{\partial A}_{kjij}$$

$$\frac{\partial E}{\partial A}_{ij} = \underbrace{\sum_k \frac{\partial E}{\partial y}_{kj} X_{ki}}_{\text{DEF OF MATMUL}}$$

$$\left(\frac{\partial E}{\partial y}\right)^T X \quad : [j, i]$$

MATRIX

BUT WE WANT i AS FIRST INDEX

TRANSPOSE!

$$\frac{\partial E}{\partial A} = \left(\left(\frac{\partial E}{\partial y}\right)^T X\right)^T$$

$$\boxed{\frac{\partial E}{\partial A} = X^T \frac{\partial E}{\partial y}}$$

FOR MORE DETAILS, CHECK:

MML BOOK, PAGE 149 (MML-BOOK.GITHUB.IO).