Derivative through matrix multiplication

① Formal derivation

Lets say we have 2 matrices $a_{(m \times p)}$, $b_{(p \times n)}$

$$z_{(m \times n)} = a \, . \, b$$

$$z_{ij} = \sum_{k=1}^{p} a_{ik} \, b_{kj} \quad \ldots \ldots (i)$$

Let say we perform certain operations on z such that

$y = f(z)$ $\{ f \rightarrow$ combination of $+$ and matmul$\}$

we perform y.backward( )

(some index k, l)

$$\frac{\partial y}{\partial a_{kl}} = \sum_{i,j} \frac{\partial y}{\partial z_{ij}} \frac{\partial z_{ij}}{\partial a_{kl}}$$

$$\frac{\partial z_{ij}}{\partial a_{kl}} = 0 \quad \{ i \neq k \} \text{ from (i)}$$

$$\Rightarrow \frac{\partial y}{\partial a_{kl}} = \sum_{j} \frac{\partial y}{\partial z_{kj}} \frac{\partial z_{kj}}{\partial a_{kl}} \quad \ldots \ldots (ii)$$

$$z_{kj} = \sum_{i=1}^{k} a_{ki} \, . \, b_{ij}$$ $\{$ same as (i) with different index $\}$

$$\frac{\partial z_{kj}}{\partial a_{kl}} = b_{lj}$$ ... substituting in (ii)

$$\frac{\partial y}{\partial a_{kl}} = \sum_j \frac{\partial y}{\partial z_{kj}} \cdot b_{lj}$$

$$= \sum_j \frac{\partial y}{\partial z_{kj}} \, b^T_{jl} \qquad \left\{ (b_{ij}) = (b^T)_{ji} \right\}$$

$\downarrow$

This equation is similar to (i) in the sense that it represents matrix multiplication ~~term~~ term.

so for all terms in a

$$\frac{\partial y}{\partial a} = G \cdot b^T \qquad \left[\begin{array}{l} \text{where } G \text{ is } \frac{\partial y}{\partial z} \\ G_{(m \times n)} \end{array} \right]$$

(gradient of a)

$$\frac{\partial y}{\partial b} = a^T \cdot G \qquad \left[\begin{array}{l} \text{wher } G \text{ is } \frac{\partial y}{\partial z} \\ G_{(m \times n)} \end{array} \right]$$

In our problem G is represented by 'gradients' parameter. (that is passed back from succeeding operation)

Then what about 'gradients == None' case. This occurs when there are no succeeding operations

i.e

$y = f(z) = z \quad \Rightarrow \quad \left[\frac{\partial y}{\partial z} = G = [1]_{m \times n} \right]$

↳ ones matrix of (m×n)

Derivative through Matrix Multiplication

② Simple informal derivation (with assumptions) [A formal derivation is also present in this document]

Let us assume we are performing matrix multiplication ~~through~~ between 2 matrices (2 x 2)

$$a = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad b = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$z = a \cdot b$$ ['.' represents matmul here]

$$z = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}$$ { z represented as }

$$z_{11} = a_{11} \times b_{11} + a_{12} \times b_{21}$$

$$z_{12} = a_{11} \times b_{12} + a_{12} \times b_{22}$$

.... So on

$$\frac{\partial z}{\partial a} = \begin{bmatrix} \partial z/\partial a_{11} & \partial z/\partial a_{12} \\ \partial z/\partial a_{21} & \partial z/\partial a_{22} \end{bmatrix}$$ { z.backward() }

$$\frac{\partial z}{\partial a_{11}} = \frac{\partial z_{11}}{\partial a_{11}} + \frac{\partial z_{12}}{\partial a_{11}} + \frac{\partial z_{21}}{\partial a_{11}} + \frac{\partial z_{22}}{\partial a_{11}}$$

$$= b_{11} + b_{12} + 0 + 0$$

$$\frac{\partial z}{\partial a_{11}} = b_{11} + b_{12}$$

similary performing for all a

$$\frac{\partial z}{\partial a} = \begin{bmatrix} b_{11} + b_{12} & b_{12} + b_{22} \\ b_{11} + b_{12} & b_{12} + b_{22} \end{bmatrix}$$

$$\frac{\partial z}{\partial a} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix} = [1]_{2\times 2} \cdot b^T$$

similary performing for b

$$\frac{\partial z}{\partial b} = a^T \cdot [1]_{2\times2}$$

This proof doesnot generalize very well, if we were to try some other examples like $(a_{(2\times3)} \cdot b_{(3\times2)})$ you will start noticing the pattern. so for a general we will get

$$z_{(m\times n)} = a_{(m\times p)} \cdot b_{(p\times n)}$$

$$\frac{\partial z}{\partial a} = [1]_{(m\times n)} \cdot b^T$$

$$\frac{\partial z}{\partial b} = a^T \cdot [1]_{m\times n}$$

You might be wondering where there random ones showup from. It comes from the fact that we are finding the derivative w.r.t to <u>z itself</u> rather than a function of z.

To generalize this further

$y = f(z)$ { f is a function made up of matmul and addition }

y.backward()

$$\frac{\partial y}{\partial a} = G \cdot b^T \qquad \left\{ \frac{\partial y}{\partial z} = G \right\} \qquad G \rightarrow m\times n$$

{ G is represented by "gradients" parameter }

$$\frac{\partial y}{\partial b} = a^T \cdot G$$

If gradients in 'None' representing that there is no succeeding operation.

This implies $f(z) = z$ (you can use eqn we derived with ones)