Exercise 3

1. Derivative of GeLU

We know: $GeLU = x \cdot P(X \le x) = x \cdot \phi(x)$ (1)

$\phi(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \xrightarrow{\mu=0,\sigma=1} \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right)$ (2)

$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

By equation (1), $\frac{\partial GeLU}{\partial x} = \phi(x) + x \cdot \phi'(x)$

By equation (2), $\phi'(x) = erf'\left(\frac{x}{\sqrt{2}}\right)$, let $z = \frac{x}{\sqrt{2}}$

By equation (3) & chain rule, $erf'(x) = \frac{\partial erf(z)}{\partial z} \cdot \frac{\partial z}{\partial x} = \frac{2}{\sqrt{\pi}} \cdot e^{-z^2} \cdot \frac{\partial z}{\partial x} = \sqrt{\frac{2}{\pi}} \cdot e^{-\frac{x^2}{2}}$

Altogether, $\frac{\partial GeLU}{\partial x} = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right) + x \cdot \sqrt{\frac{2}{\pi}} \cdot e^{-\frac{x^2}{2}}$

2. Derivative of Leaky ReLU.

if $x > 0$, $\frac{\partial L\text{-}ReLU}{\partial x} = 1$

if $x = 0$, define $\frac{\partial L\text{-}ReLU}{\partial x}$ as 1.

if $x < 0$, $\frac{\partial L\text{-}ReLU}{\partial x} = \frac{\partial \alpha x}{\partial x} = \alpha$, where α is the constant defined in L-ReLU.

3. Prove the backprop through einsum.

We know: $A \in R^{N \times M}$, $B \in R^{M \times K}$, $C = einsum("\overset{=u}{ij}, \overset{=v}{jk} \to \overset{=w}{ik}", A, B)$ (1)

equa. (1) implies: $C_{ik} = \sum_{j=1}^{M} a_{ij} b_{jk}$ (2)

$\frac{\partial L}{\partial a_{ij}} = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\partial L}{\partial C_{ik}} \cdot \frac{\partial C_{ik}}{\partial a_{ij}} = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\partial L}{\partial C_{ik}} \cdot \frac{\sum_{j=1}^{M} \partial a_{ij} b_{jk}}{\partial a_{ij}}$ (3)

Only when $i = \tilde{i}$, $j = \tilde{j}$, the derivative ($\frac{\partial a_{ij} b_{jk}}{\partial a_{ij}}$) w.r.t. a_{ij} is non-zero, hence

$D_{(ij)} := \frac{\partial L}{\partial a_{ij}} = \sum_{k=1}^{K} \frac{\partial L}{\partial C_{ik}} \cdot b_{jk}$

$\overset{u}{\|}$ $\quad \because C\text{-}grad_{(ik)}$ $\overset{=w}{\quad} \overset{=v}{\quad} \overset{=u}{\quad}$

which implies $D = einsum("ik, jk \to ij", C\text{-}grad, B)$