RMSNorm Backward Pass Derivation for one row

Forward Pass

Given input $x \in \mathbb{R}^{1\times N}$, RMSNorm is defined as:

$$y = \text{RMSNorm}(x) = \frac{x}{\text{RMS}(x)} \odot g \tag{1}$$

where:

$$\text{RMS}(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 + \epsilon} \tag{2}$$

Backward Pass

Given gradient $dy = \frac{\partial L}{\partial y} \in \mathbb{R}^{1\times N}$:

$$dx = dy@J \tag{3}$$

The expanded matrix multiplication becomes:

$$\begin{bmatrix} dx_1 & dx_2 & \cdots & dx_N \end{bmatrix} = \begin{bmatrix} dy_1 & dy_2 & \cdots & dy_N \end{bmatrix} \begin{bmatrix} - & \frac{\partial y_1}{\partial x} & - \\ - & \frac{\partial y_2}{\partial x} & - \\ \vdots & \vdots & \vdots \\ - & \frac{\partial y_N}{\partial x} & - \end{bmatrix} \tag{4}$$

$$= \begin{bmatrix} dy_1 & dy_2 & \cdots & dy_N \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_N} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_N}{\partial x_1} & \frac{\partial y_N}{\partial x_2} & \cdots & \frac{\partial y_N}{\partial x_N} \end{bmatrix} \tag{5}$$

Jacobian Derivation

Let $y_j = \frac{x_j}{r}$ where $r = \text{RMS}(x)$.

Step 1: Apply the product rule:

$$\frac{\partial y_j}{\partial x_k} = \frac{\partial x_j}{\partial x_k} \cdot r^{-1} + x_j \cdot \frac{\partial (r^{-1})}{\partial x_k} \tag{6}$$

Step 2: Compute $\frac{\partial r}{\partial x_k}$:

$$\frac{\partial r}{\partial x_k} = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-1/2} \cdot \frac{2x_k}{N} = \frac{x_k}{N \cdot r} \tag{7}$$

Step 3: Compute $\frac{\partial(r^{-1})}{\partial x_k}$:

$$\frac{\partial(r^{-1})}{\partial x_k} = -r^{-2} \cdot \frac{x_k}{N \cdot r} = -\frac{x_k}{N \cdot r^3} \tag{8}$$

Step 4: Final Jacobian:

$$J_{jk} = \frac{\partial y_j}{\partial x_k} = \frac{\delta_{jk}}{r} - \frac{x_j x_k}{N \cdot r^3} = \frac{1}{r}\left(\delta_{jk} - \frac{x_j x_k}{N \cdot r^2}\right) \tag{9}$$

In matrix form:

$$J_{(N \times N)} = \frac{1}{r}\left(I - \frac{x^T x}{N \cdot r^2}\right) \tag{10}$$

- x is a row vector here. so the shape of $x^T x$ is (N, N).

Vector form:

$$dx = dy \cdot J = \frac{1}{r}\left(dy - \frac{dy \cdot x^T \cdot x}{Nr^2}\right) = \frac{1}{r}\left(dy - \underbrace{(dy \cdot x^T)}_{scalar}\frac{1}{Nr^2}x\right) \tag{11}$$

With scale parameter g:

$$dx = (g \odot dy) \cdot J = \frac{1}{r}\left((g \odot dy) - \underbrace{((g \odot dy) \cdot x^T)}_{scalar}\frac{1}{Nr^2}x\right) \tag{12}$$

$$= \frac{1}{r}\left((g \odot dy) - \underbrace{\frac{(g \odot dy) \cdot x^T}{N \cdot r^2}}_{scalar} \odot x\right) \tag{13}$$

$$= \frac{1}{r}\left((g \odot dy) - \underbrace{\frac{(g \odot dy) \cdot \hat{x}^T}{N}}_{scalar} \odot \hat{x}\right), Let\ \hat{x} = \frac{x}{r} \tag{14}$$

$$\tag{15}$$