softmax:

$$\hat{y}_i = \frac{e^{z_i}}{\sum e^{z_j}}$$

cross-entropy loss:

$$L = -\sum y_i \cdot \log(\hat{y}_i)$$

$$L = -\log(\hat{y}_k)$$ k is true, so $y_k = 1$

$$L = -\log\left(\frac{e^{z_k}}{\sum e^{z_j}}\right)$$

with $\log\left(\frac{a}{b}\right) = \log a - \log b$

$$L = -\log(e^{z_k}) + \log\left(\sum e^{z_j}\right)$$

$$= -z_k + \log\left(\sum e^{z_j}\right)$$

$$\frac{dL}{dz_i} = \frac{d}{dz_i} - z_i + \frac{d}{dz_i} \log\left(\sum e^{z_j}\right) = -1 + \hat{y}_i \quad \text{for true}$$

$$= \hat{y}_i \quad \text{for false}$$

$$\frac{d}{dz_i} - z_i = -1 \quad \text{for true}$$

$$= 0 \quad \text{for false}$$

$$\frac{d}{dz_i} \log\left(\sum e^{z_j}\right)$$

chain rule:

$$= \frac{1}{\sum e^{z_j}} \cdot \frac{d}{dz_i} \sum e^{z_j} = \frac{e^{z_i}}{\sum e^{z_j}}$$

layernorm: H: d_{model}

$$y_i = \hat{x}_i \cdot w_i + b_i \qquad \mu = \frac{\sum x_i}{H} \qquad \sigma^2 = \frac{\sum (x_j - \mu)^2}{H} \qquad \hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2}}$$

σ^2: variance
$\sqrt{\sigma^2}$: standard dev

$$\frac{dL}{db_i} = \frac{dL}{dy_i} \cdot \frac{dy_i}{db_i} = \frac{dL}{dy_i} \cdot 1$$

$$\frac{dL}{dw_i} = \frac{dL}{dy_i} \cdot \frac{dy_i}{dw_i} = \frac{dL}{dy_i} \cdot \hat{x}_i$$

$$\frac{dL}{d\hat{x}_i} = \frac{dL}{dy_i} \cdot \frac{dy_i}{d\hat{x}_i} = \frac{dL}{dy_i} \cdot w_i$$

$$\frac{dL}{dx_i} = \frac{dL}{d\hat{x}_i} \cdot \frac{d\hat{x}_i}{dx_i}$$

$$+ \sum_j \left(\frac{dL}{d\hat{x}_j} \cdot \frac{d\hat{x}_j}{d\mu}\right) \cdot \frac{d\mu}{dx_i}$$

$$+ \sum_j \left(\frac{dL}{d\hat{x}_j} \cdot \frac{d\hat{x}_j}{d\sigma^2}\right) \cdot \frac{d\sigma^2}{dx_i}$$

$$+ \sum_j \left(\frac{dL}{d\hat{x}_j} \cdot \frac{d\hat{x}_j}{d\sigma^2}\right) \cdot \frac{d\sigma^2}{d\mu} \cdot \frac{d\mu}{dx_i}$$

$$= \frac{dL}{d\hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2}}$$

$$- \sum_j \left(\frac{dL}{d\hat{x}_j}\right) \cdot \frac{1}{H} \cdot \frac{1}{\sqrt{\sigma^2}}$$

$$- \sum_j \left(\frac{dL}{d\hat{x}_j} \cdot \hat{x}_j\right) \cdot \frac{1}{H} \cdot \hat{x}_i \cdot \frac{1}{\sqrt{\sigma^2}}$$

$$\frac{d\hat{x}_i}{dx_i} = \frac{1}{\sqrt{\sigma^2}} \qquad \frac{d\hat{x}_i}{d\mu} = -\frac{1}{\sqrt{\sigma^2}}$$

$$\hat{x}_i = (x_i - \mu) \cdot (\sigma^2)^{-\frac{1}{2}}$$

$$\frac{d\hat{x}_i}{d\sigma^2} = -\frac{1}{2}(x_i - \mu) \cdot (\sigma^2)^{-\frac{3}{2}}$$

$$\frac{d\mu}{dx_i} = \frac{1}{H} \qquad \frac{d\sigma^2}{dx_i} = \frac{d}{dx_i} \frac{(x_i - \mu)^2}{H} = \frac{x_i^2 - 2x_i\mu + \mu^2}{H}$$

$$= \frac{2x_i - 2\mu}{H}$$

$$= \frac{2(x_i - \mu)}{H}$$

$$\frac{d\sigma^2}{d\mu} = \frac{(x_0 - \mu)^2}{H} + \frac{(x_1 - \mu)^2}{H} + \dots$$

$$= \frac{-2x_0 + 2\mu}{H} + \frac{-2x_1 + 2\mu}{H} + \dots$$

$$= \frac{\sum_j -2(x_j - \mu)}{H} = -\frac{2}{H} \sum_j (x_j - \mu) = 0$$

sum of $x_j - \mu$ is zero by definition

$$\sum_j \left(\frac{dL}{d\hat{x}_j} \cdot \frac{1}{\cancel{2}} (x_j - \mu) \cdot \frac{1}{\sqrt{\sigma^2}}\right) \cdot \frac{1}{\sqrt{\sigma^2}} \cdot \frac{1}{\sqrt{\sigma^2}} \cdot \cancel{2} \frac{x_i - \mu}{H}$$

softmax:

$$y_i = \frac{e^{x_i}}{\sum e^{x_j}} \qquad \frac{dy_k}{dx_i} = \frac{d}{dx_i} \frac{e^{x_k}}{\sum_j e^{x_j}}$$

quotient rule $\frac{f(x)}{g(x)} = \frac{f' \cdot g - f \cdot g'}{g^2}$

for $i = k$:

$$\frac{dy_k}{dx_k} = \frac{e^{x_k} \cdot \sum_j e^{x_j} - e^{x_k} \cdot e^{x_k}}{(\sum_j e^{x_j})^2}$$

$$= \frac{e^{x_k}(\sum_j e^{x_j} - e^{x_k})}{(\sum_j e^{x_j})^2}$$

$$= y_k \cdot \frac{\Sigma - e^{x_k}}{\Sigma}$$

$$= y_k \cdot (1 - y_k)$$

for $i \neq k$:

$$\frac{dy_k}{dx_i} = \frac{0 \cdot \sum_j e^{x_j} - e^{x_k} \cdot e^{x_i}}{(\sum_j e^{x_j})^2}$$

$$= -y_k \cdot y_i$$

$$\frac{dL}{dx_i} = \sum_k \frac{dL}{dy_k} \cdot \frac{dy_k}{dx_i}$$

$$\frac{dL}{dx_0} = \frac{dL}{dy_0} \cdot y_0 \cdot (1-y_0) = \frac{dL}{dy_0} \cdot y_0 + \frac{dL}{dy_0} \cdot -y_0 \cdot y_0 = \frac{dL}{dy_0} \cdot y_0 - \sum_j \left(\frac{dL}{y_j} \cdot y_j \right) \cdot y_0$$

$$+ \frac{dL}{dy_1} \cdot -y_1 \cdot y_0 \qquad + \frac{dL}{dy_1} \cdot -y_1 \cdot y_0$$

$$+ \frac{dL}{dy_2} \cdot -y_2 \cdot y_0 \qquad + \dots$$

$$+ \dots$$

$$\frac{dL}{dx_i} = y_i \cdot \left[\frac{dL}{dy_0} - \sum_j \left(\frac{dL}{dy_j} \cdot y_j \right) \right]$$

expressed as Jacobian:

$$J = \begin{bmatrix} y_0 - y_0^2 & -y_0 \cdot y_1 & \cdots \\ -y_1 \cdot y_0 & y_1 - y_1^2 & \\ \vdots & & \end{bmatrix} = \begin{bmatrix} y_0 & 0 & \cdots \\ 0 & y_1 & \\ \vdots & & \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix} \times [y_0, y_1, y_2] \qquad J_{k,i} = \frac{dy_k}{dx_i}$$

gelu:

$y = x \cdot cdf(x)$

with cdf as the cumulative distribution function of the standard normal distribution $\mu = 0, \sigma^2 = 1$

$$cdf(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \quad \text{implemented as:} \quad cdf(x) = \frac{1}{2} \cdot \left(1 + erf\left(\frac{x}{\sqrt{2}} \right) \right)$$

$$\frac{dy}{dx} = \frac{d}{dx} x \cdot cdf(x) + x \cdot \frac{d}{dx} cdf(x)$$

$$\text{since } \frac{d}{dx} \int_{-\infty}^{x} f(t)\, dt = f(x)$$

$$\frac{dy}{dx} = cdf(x) + x \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x^2}{2}}$$

two ways of matmul used in the code:

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 4 \\ 0 & 0 & 3 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 22 \\ 12 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 2 \\ 6 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 16 \\ 12 \end{bmatrix}$$

Which one is faster depends on data layout