Linear

One sample: $y = w_1 x_1 + w_2 x_2 + w_3 x_3 + b \quad y \to \mathcal{L}$

$$y = \underset{1\times 3}{[x_1\ x_2\ x_3]} \underset{3\times 1}{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}} + \underset{1\times 1}{[b]}$$

① $$\frac{d\mathcal{L}}{dw} = \frac{d\mathcal{L}}{dy} \cdot \frac{dy}{dw} = \underset{[3\times 1]}{\begin{bmatrix} dy/dw_1 \\ dy/dw_2 \\ dy/dw_3 \end{bmatrix}} \underset{[1\times 1]}{\frac{d\mathcal{L}}{dy}}$$

$$= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \frac{d\mathcal{L}}{dy} = x^T \frac{d\mathcal{L}}{dy}$$

a constant 1×1 matrix side dowd matt

② $$\frac{d\mathcal{L}}{db} = \frac{d\mathcal{L}}{dy} \cdot \overbrace{\frac{dy}{db}}^{1} = \frac{d\mathcal{L}}{dy}$$

③ $$\underset{[1\times 1]}{\frac{d\mathcal{L}}{dx}} = \underset{[1\times 1]}{\frac{d\mathcal{L}}{dy}} \cdot \underset{[1\times 3]}{\frac{dy}{dx}} = \frac{d\mathcal{L}}{dy} [w_1\ w_2\ w_3] = \frac{d\mathcal{L}}{dy} W^T$$

Batch Dimensions

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \underset{2\times 3}{\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix}} \underset{3\times 1}{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}} + [b]$$

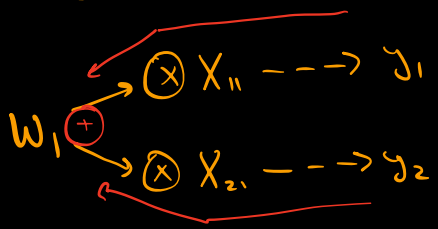
broadcasting

$$\frac{dz}{dw} = \begin{bmatrix} dz/dw_1 \\ dz/dw_2 \\ dz/dw_3 \end{bmatrix}$$

$[2\times1]$ $[1\times1]$

Which y? y_1 or y_2?

① $$\frac{dz}{dw_1} = \frac{dz}{dy} \cdot \frac{dy}{dw_1}$$



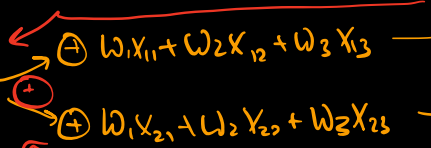
$$= \sum_{i=1}^{N} \frac{dz}{dy_i} \cdot \frac{dy_i}{dw_1} = \frac{dz}{dy_1} \cdot X_{11} + \frac{dz}{dy_2} \cdot X_{21}$$

$$\frac{dy}{dw} = \begin{bmatrix} \frac{dL}{dy_1} \cdot X_{11} + \frac{dL}{dy_2} X_{21} \\ \frac{dL}{dy_1} \cdot X_{12} + \frac{dL}{dy_2} X_{22} \\ \frac{dL}{dy_1} \cdot X_{13} + \frac{dL}{dy_2} X_{23} \end{bmatrix}$$

Sum over batch is contained in the matmul!

$$= \begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ X_{13} & X_{23} \end{bmatrix} \begin{bmatrix} dL/dy_1 \\ dL/dy_2 \end{bmatrix} = \boxed{X^T \frac{dL}{dy}}$$

② $$\frac{dL}{db}$$



but $\frac{dz}{db} = \frac{dL}{db} \cdot \frac{dy}{db} = \frac{dL}{db} \cdot 1$ so $\boxed{\sum_{i=1}^{N} \frac{dL}{dy_i}}$

③ $\frac{dL}{dX} = \frac{dL}{dy} W^T$ ← same as before, all the batch grad contributions contained inside sum.

Softmax (Slow)

More complex as its a vector function:

$$\vec{O} = [O_1 \; O_2 \; O_3 \cdots O_N]$$

$$\text{Softmax}(\vec{O}) = [S_1 \; S_2 \; S_3 \cdots S_N]$$

Where $S_i = \frac{e^{O_i}}{\sum_{k=1}^{N} e^{O_k}}$

interdependency btwn some i and j so two cases.

if $i=j$

$$\frac{dS_i}{dO_j} = \frac{\left[\sum_{k=1}^{N} e^{O_k}\right] e^{O_i} - e^{O_i} e^{O_j}}{\left(\sum_{k=1}^{N} e^{O_k}\right)^2}$$

$$= \frac{e^{O_i}\left[\sum_{k=1}^{N} e^{O_k} - e^{O_j}\right]}{\left(\sum_{k=1}^{N} e^{O_k}\right)^2}$$

$$= \frac{e^{O_i}}{\sum_{k=1}^{N} e^{O_k}} \cdot \frac{\sum_{k=1}^{N} e^{O_k} - e^{O_j}}{\sum_{k=1}^{N} e^{O_k}}$$

$$S_i \left(1 - S_j\right)$$

if $i \neq j$

$$\frac{dS_i}{dO_j} = \frac{\left[\sum_{k=1}^{N} e^{O_k}\right] 0 - e^{O_i} e^{O_j}}{\left[\sum_{k=1}^{N} e^{O_k}\right]^2}$$

$$= - \frac{e^{O_i}}{\sum_{k=1}^{N} e^{O_k}} \frac{e^{O_j}}{\sum_{k=1}^{N} e^{O_k}} = - S_i S_j$$

$S_1 - S_1 S_1$ $\qquad$ $S_2 - S_2 S_2$

$$J = \begin{bmatrix} S_1(1-S_1) & -S_1 S_2 & -S_1 S_3 \\ -S_2 S_1 & S_2(1-S_2) & -S_2 S_3 \\ -S_3 S_1 & -S_3 S_2 & S_3(1-S_3) \end{bmatrix}$$

Numerical Stability:

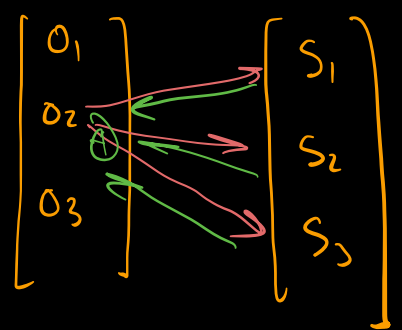
$$\frac{e^{x-c}}{\sum_i e^{x_i - c}} = \frac{e^x / e^c}{(\sum_i e^{x_i})/e^c}$$

Per sample jacobian!

if N is large this is a massive N^2 matrix!

Can we do better?

What we want: $\frac{dL}{dO_j}$ and O_j has grad contributions from all S_j



$$\frac{dL}{dO_j} = \sum_i \frac{dL}{dS_i} \cdot \frac{dS_i}{dO_j}$$

(we already computed this!)

$$= \sum_i \frac{dL}{dS_i} \left[S_i \left[\delta_{ij} - S_j \right] \right]$$

Split sum into 2 cases.

Where $\delta_{ij} = 1$ if $i=j$ else 0.

$$= \frac{dL}{ds_j}\left[s_j(1-s_j)\right] + \sum_{\substack{i \\ i\neq j}} \frac{dL}{ds_i}\left(-s_i s_j\right)$$

$$= \frac{dL}{ds_j} s_j - \frac{dL}{ds_j} s_j s_j - \sum_{\substack{i \\ i\neq j}} \frac{dL}{ds_i}\left(-s_i s_j\right)$$

$$= \frac{dL}{ds_j} s_j - \sum_{i} \frac{dL}{ds}\left(s_i s_j\right)$$

Sum doesn't depend on s_j

$$= s_j\left[\frac{dL}{ds_j} - \sum_{i} \overbrace{\frac{dL}{ds_i} s_i}^{\text{dot prod.}}\right]$$

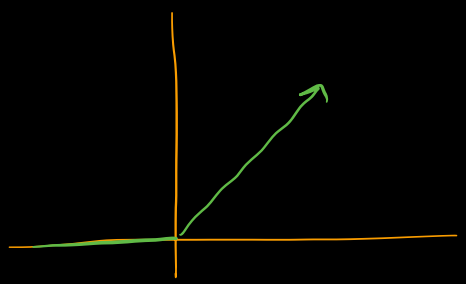
↓

Vector notation

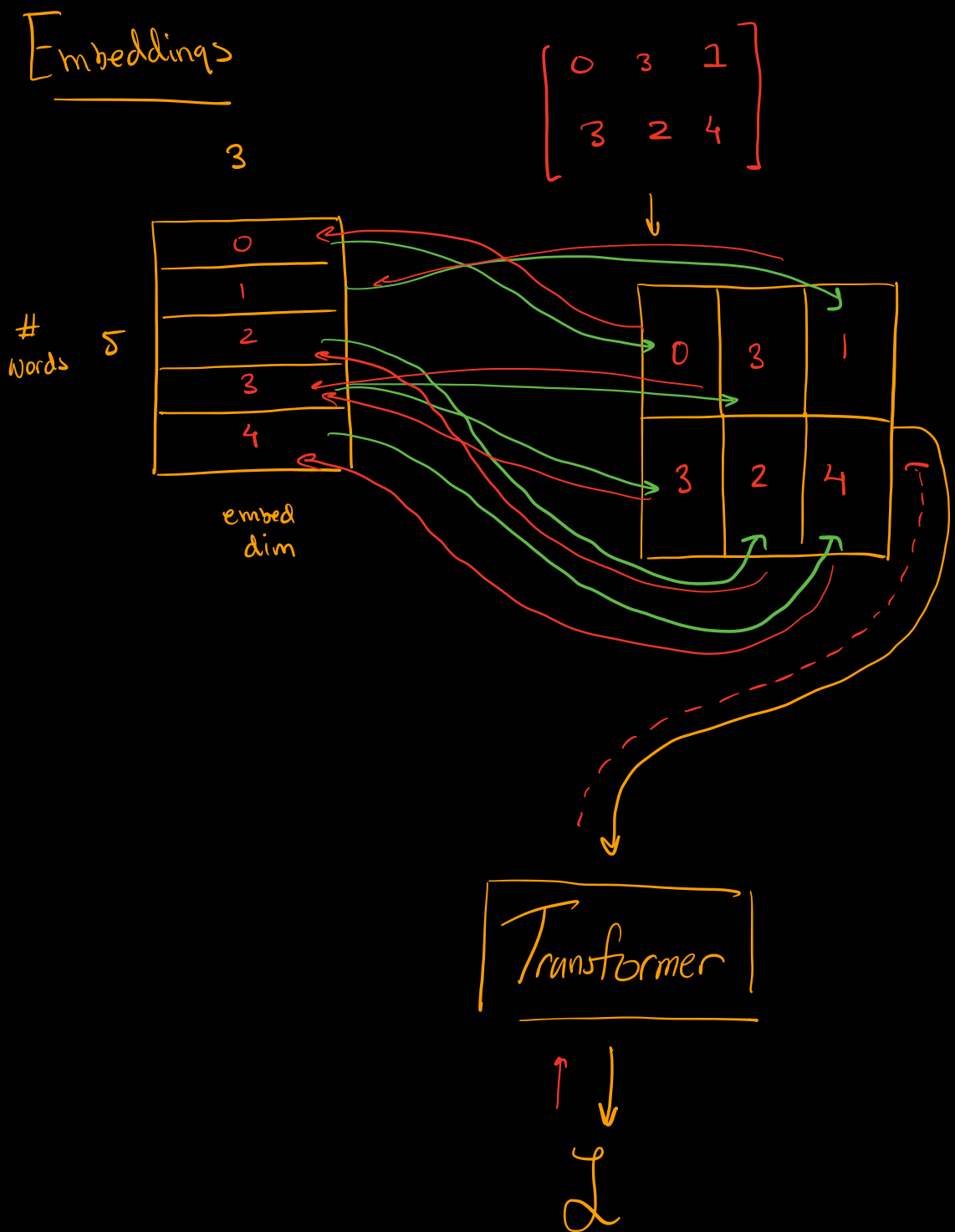
$$S \cdot \left[\frac{dL}{ds} - \left[\frac{dL}{ds} \cdot S\right]\right]$$

Relu

$$\begin{bmatrix} -2 \\ 3 \\ 8 \end{bmatrix} \rightarrow \text{Relu}(x) \rightarrow \begin{bmatrix} 0 \\ 3 \\ 8 \end{bmatrix}$$



$$\frac{dy}{dx} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else} \end{cases}$$


Embeddings
3
words
5
0
1
2
3
4
embed dim
0 3 1
3 2 4
Transformer
L

Layer Norm

Data: $B \times S \times E$

normalize along last dimension E.

$$\mu_{b,s} = \frac{1}{E} \sum_{e=1}^{E} x_{b,s,e}$$ ← w/i x interdependency

$$Var_{b,s} = \frac{1}{E} \sum_{e=1}^{E} (x_{b,s,e} - \mu_{b,s})^2$$

$$\hat{x}_{b,s,e} = \frac{x_{b,s,e} - \mu_{b,s}}{\sqrt{var_{b,s} + \varepsilon}}$$

$$y_{b,s,e} = \gamma_e \cdot \hat{x}_{b,s,e} + \beta_e$$

γ_e: learnable scale param

β_e: learnable shift param.

given upstream grad $\frac{dL}{dy_{b,s,e}}$ we need

$$\frac{dL}{d\gamma_e}, \frac{dL}{db_e}, \frac{dL}{dx_{b,s,e}}$$

Remember, our B, S indexes are independent, ∴ the grad are accumulated. For simplicity I will ignore them.

$$\mu = \frac{1}{E}\sum_{k=1}^{E} X_k$$

$$u \triangleq X_k - \mu \leftarrow \text{mean centered.}$$

$$\sigma^2 = \frac{1}{E}\sum_{k=1}^{E} u_k^2$$

$$S = \sqrt{\sigma^2 + \varepsilon}$$

$$\hat{X}_j = \frac{u_j}{S} = \frac{X_j - \mu}{S}$$

($X_j - \mu$ ← has X in it; S ← has X in it)

$$y_j = \gamma_j \hat{X}_j + \beta_j$$

$$\nabla_j = \frac{dL}{dy_j} \leftarrow \text{upstream gradient.}$$

$$\frac{dL}{dX_i} = \sum_{j=1}^{E} \frac{dL}{dy_j} \cdot \frac{dy_j}{dX_i} \leftarrow \text{Sum up all those contributions.}$$

$$= \sum_{j=1}^{E} \nabla_j \underbrace{\frac{d y_j}{d \hat{x}_j}}_{\gamma_j} \cdot \frac{d \hat{x}_j}{d x_i}$$

$$= \sum_{j=1}^{E} \nabla_j \gamma_j \underbrace{\frac{d \hat{x}_j}{d x_i}}_{\text{Solve this:}}$$

$$\mu = \frac{1}{E} \sum_{k=1}^{E} x_k \rightarrow \boxed{\frac{d\mu}{dx_i} = \frac{1}{E}}$$

$$u_j = x_j - \mu \rightarrow \boxed{\frac{du_j}{dx_i} = \delta_{ij} - \frac{1}{E}}$$

δ_{ij}: 1 if $i=j$ else 0.

$$\sigma^2 = \frac{1}{E} \sum_{k=1}^{E} u_k^2$$

$$\frac{d\sigma^2}{dx_i} = \frac{1}{E} \sum_{k=1}^{E} 2 u_k \cdot \frac{du_k}{dx_i}$$

$$= \frac{2}{E} \sum_{k=1}^{E} u_k \left[\delta_{ik} - \frac{1}{E} \right]$$

$$= \frac{2}{E} \left[\sum_{k=1}^{E} u_k \delta_{ik} - \sum_{k=1}^{E} u_k \frac{1}{E} \right]$$

δ_{ik}: 0 if $i \neq k$

$$= \frac{2}{E} \left[u_i - \frac{1}{E} \sum_{k=1}^{E} u_k \right]$$

$$\frac{1}{E} \sum_{k=1}^{E} u_k = \frac{1}{E} \sum_{k=1}^{E} x_k - \mu = 0$$

mean centered data. So mean of mean centered is 0!

$$\therefore \boxed{\frac{d\sigma^2}{dx_i} = \frac{2}{E} u_i}$$

$$S = \sqrt{\sigma^2 + \varepsilon} = (\sigma^2 + \varepsilon)^{1/2}$$

$$\frac{ds}{dx_i} = \frac{1}{2}(\sigma^2 + \varepsilon)^{-\frac{1}{2}} \cdot \frac{d(\sigma^2 + \varepsilon)}{dx_i}$$

(constant)

$$= \frac{1}{2\sqrt{\sigma^2 + \varepsilon}} \frac{2}{E} u_i$$

$$= \frac{1}{S} \frac{1}{E} u_i$$

$$\therefore \boxed{\frac{ds}{dx_i} = \frac{u_i}{SE}}$$

$$\hat{x}_j = \frac{u_j}{S} = u_j \left(\frac{1}{S}\right)$$

(Product rule)

$$\frac{d\hat{x}_j}{dx_i} = \underbrace{u_j \cdot \frac{d(1/S)}{dx_i}}_{①} + \underbrace{\frac{du_j}{dx_i} \frac{1}{S}}_{②}$$

① $$u_j \frac{d\left(\frac{1}{s}\right)}{dx_i} = u_j \frac{d}{dx_i} s^{-1} = -s^{-2} \frac{ds}{dx_i} u_j$$

$$= -\frac{1}{s^2} u_j \frac{u_i}{sE} = -\frac{u_i u_j}{s^3 E}$$

② $$\frac{du_j}{dx_i} \frac{1}{s} = \left(\delta_{ij} - \frac{1}{E}\right)\left(\frac{1}{s}\right)$$

$$\therefore \frac{d\hat{x}_j}{dx_i} = -\frac{u_i}{s^3 E} + \left(\delta_{ij} - \frac{1}{E}\right)\left(\frac{1}{s}\right)$$

$$= \frac{1}{s}\left[\delta_{ij} - \frac{1}{E} - \frac{u_i u_j}{s^2 E}\right] \qquad \hat{x} = \frac{u}{s}$$

$$= \frac{1}{s}\left[\delta_{ij} - \frac{1}{E} - \frac{u_i}{s}\frac{u_j}{s}\frac{1}{E}\right]$$

$$= \frac{1}{s}\left[\delta_{ij} - \frac{1}{E} - \frac{\hat{x}_i \hat{x}_j}{E}\right]$$

Now remember that: we just solved this.

$$\frac{dL}{dx_i} = \sum_{j=1}^{E} \nabla_j \gamma_j \frac{\delta \hat{x}_j}{\delta x_i}$$

$$= \sum_{j=1}^{E} \nabla_j \gamma_j \frac{1}{s}\left[\delta_{ij} - \frac{1}{E} - \frac{\hat{x}_i \hat{x}_j}{E}\right]$$

$$= \frac{1}{s}\left[\sum_{j=1}^{E} \nabla_j \gamma_j \delta_{ij} - \frac{1}{E}\sum_{j=1}^{E} \nabla_j \gamma_j - \frac{1}{E}\sum_{j=1}^{E} \nabla_j \gamma_j \hat{x}_i \hat{x}_j\right]$$

($\delta_{ij} = 1$ if $i=j$)

$$= \frac{1}{s}\left[\nabla_i \gamma_i - \frac{1}{E}\sum_{j=1}^{E} \nabla_j \gamma_j - \frac{\hat{x}_i}{E}\sum_{j=1}^{E} \nabla_j \gamma_j \hat{x}_j\right]$$

Now remember

$$\nabla = \frac{dL}{dy} \quad \text{and} \quad \frac{dy}{d\hat{x}} = \gamma$$

So for simplicity, our upstream grad to $\hat{x}$ (not just y) is:

$$\frac{dL}{dy} \cdot \frac{dy}{d\hat{x}} = \nabla\gamma \triangleq \tilde{\nabla}$$ ← let's just give it a new symbol

$$= \frac{1}{s}\left[\tilde{\nabla} - \frac{1}{E}\sum_{j=1}^{E}\tilde{\nabla} - \frac{\hat{x}_i}{E}\sum_{j=1}^{E}\tilde{\nabla}\hat{x}_j\right]$$

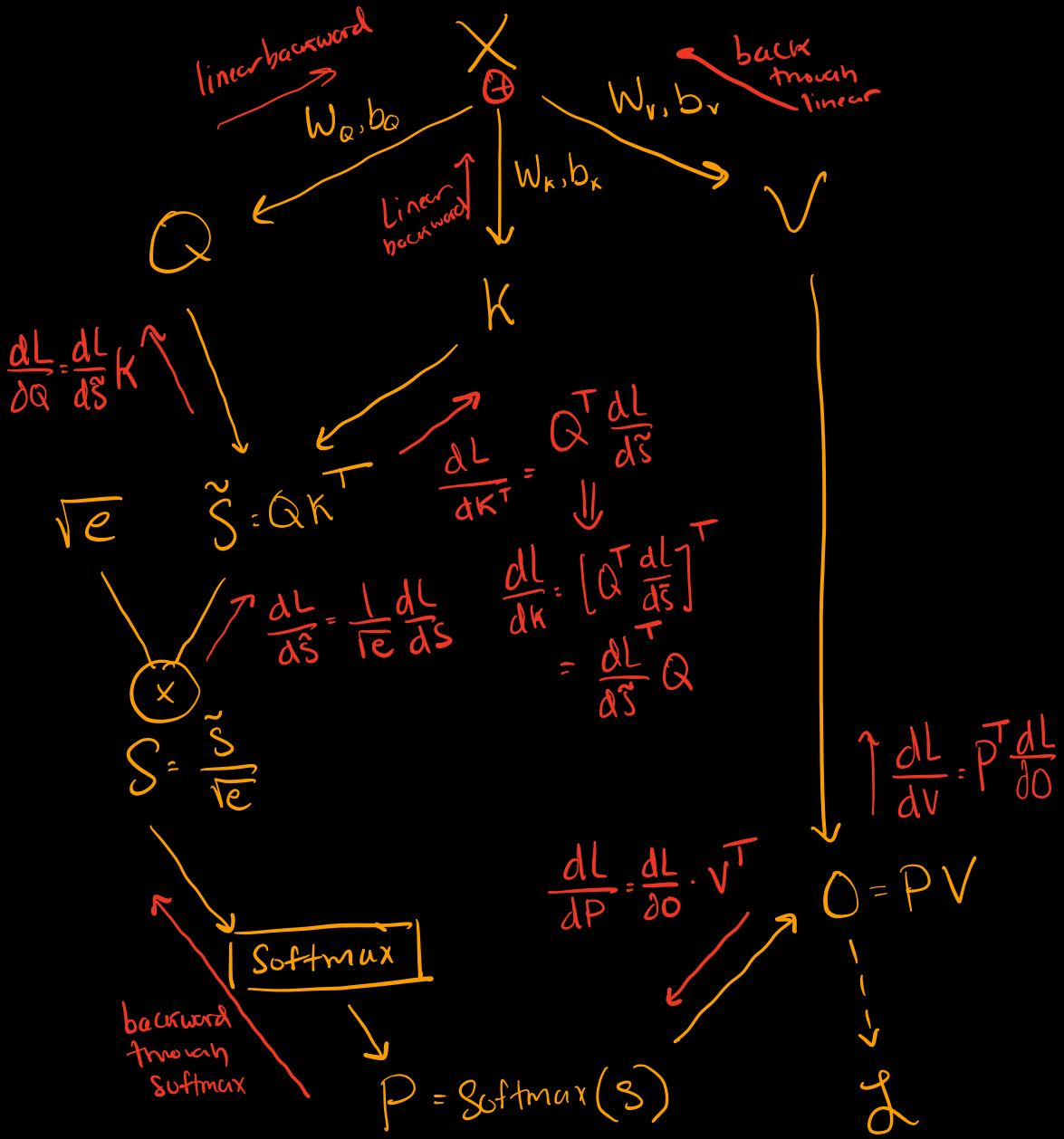
dot product

$$= \frac{1}{\sqrt{\sigma^2+\varepsilon}}\left[\tilde{\nabla} - \text{mean}(\tilde{\nabla}) - \hat{x}\cdot\text{mean}(\tilde{\nabla}\cdot\hat{x})\right]$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{\beta,s}^{1} \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial \beta_j} = \boxed{\sum_{\beta,s}^{1} \nabla}$$

$$\frac{\partial L}{\partial \gamma_j} = \sum_{\beta,s}^{1} \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial \gamma_j} = \boxed{\sum_{\beta,s}^{1} \nabla \hat{x}_j}$$

Multihead Attention



Cross Entropy Loss (For OHE Labels)

$$\mathcal{L}(P, y) = -\sum_{i=1}^{n} y_i \log(P_i)$$

P is a probability (softmaxed)

y is one hot:

$$[\,0 \quad 0 \quad 1 \quad 0 \quad 0\,]$$

$$\mathcal{L}(\,, y) = -\Big[\, 0\log(P_1) + 0\log(P_2) + 1\log(P_3) + 0\log(P_4) + 0\log(P_5) \,\Big]$$

$$= -\log(P_3)$$

$$\therefore \frac{d\mathcal{L}}{dP_i} = -\frac{y_i}{P_i} = \begin{cases} 0 & \text{if } i \text{ is not correct} \\ -\frac{1}{P_i} & \text{if } i \text{ is correct.} \end{cases}$$

logits → Softmax → loss

Problem. Previous derivation expects already softmaxed outputs. This means in the backward pass we have to backprop through softmax.

Can we avoid that?

$$CE = -\sum_j y_j \log(P_j)$$

$$\text{where } P_j = \frac{e^{x_j}}{\sum_i e^{x_i}}$$

but y is OHE so

$$CE = -\log(P_{label})$$

$$= -\log\left[\frac{e^{x_{label}}}{\sum_i e^{x_i}}\right]$$

$$= -\left[\log\left(e^{x_{label}}\right) - \log\sum_i e^{x_i}\right]$$

$$= \log \sum_i e^{x_i} - \log\left(e^{x_{label}}\right)$$

$$= \log \sum_i e^{x_i} - x_{label}$$

$$\frac{dL}{dx_j}\left[\log \sum_i e^{x_i} - x_{label}\right]$$

$$= \frac{1}{\sum_i e^{x_i}} \frac{d}{dx_j} \sum_i e^{x_i} - \begin{cases} 1 & \text{if } j = \text{label} \\ 0 & \text{else} \end{cases}$$

$$= \frac{e^{x_j}}{\sum_i e^{x_i}} - \begin{cases} 1 & \text{if } j = \text{label} \\ 0 & \text{else} \end{cases}$$

Our grad uses our probs!

$$= P_j - \delta_j \quad \text{where } \delta_j = \begin{cases} 1 & \text{if } j = \text{label} \\ 0 & \text{else.} \end{cases}$$

logits → loss