Deriving Self-Attention Head (Q_x, K_x, V_x, Emb$_{in}$)

$$\frac{\partial L}{\partial V} = \left(M^T @ \frac{\partial L}{\partial out}\right) . \text{Sum}(0)$$

$$\frac{\partial L}{\partial M_2} = \frac{\partial L}{\partial out} @ V^T$$

$$\frac{\partial L}{\partial M_1} = \frac{-\text{softmax}_i \sum_n \text{softmax}_n}{\text{softmax}_i + \text{softmax}_i}$$ row$_n$ ↑ mult by $\frac{\partial L}{\partial M_2}$

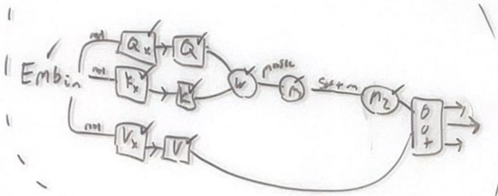
$\frac{\partial L}{\partial w}$ → Sel/'stop' term goes

$$\frac{\partial L}{\partial K} = \frac{\partial L}{\partial w} \cdot \frac{\partial w}{\partial Q} = dout^T @ Q$$

$$\frac{\partial L}{\partial Q} = \frac{\partial L}{\partial w} \cdot \frac{\partial w}{\partial K} = dout @ K$$

$Q @ K^T = w$

1 head Self-Attention



$$\frac{\partial L}{\partial Q_x} = \frac{\partial L}{\partial Q} \cdot \frac{\partial Q}{\partial Q_x}$$ → Same for K_x, V_x

$Q = Emb @ Q_x$

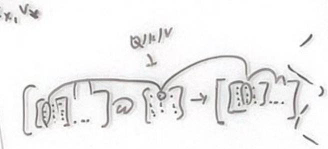
• For this, EL_{IJ} affects jth col out by dot w/ Emb Ith col

$$\frac{\partial L}{\partial Q_x} = Emb^T @ \frac{\partial L}{\partial Q}$$

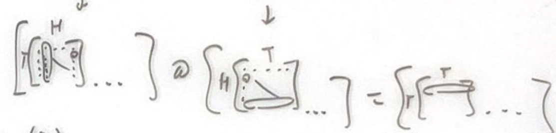
$$\frac{\partial L}{\partial K_x} = Emb^T @ \frac{\partial L}{\partial K}$$

$$\frac{\partial L}{\partial V_x} = Emb^T @ \frac{\partial L}{\partial V}$$

→ Affects all Batches, so sum Dim=0



Q ↓ K^T ↓



(B)

• $EL_{I,J}$ of K is $(EL_{J,I})$ affects Jth col of out by Ith col of Q

$$\frac{\partial L}{\partial EL_{I,J}} = \sum_x Q_{xJ} \cdot \frac{\partial L}{\partial out_{xI}}$$

Dot jth col Q w/ ith col $\frac{\partial L}{\partial out}$

$dout^T @ Q$

↓

Each $EL_{I,J}$ of Q affects Ith row of $\frac{\partial L}{\partial out}$ by jth row

$$\sum_x K^T_{Jx} \cdot \frac{\partial L}{\partial out_{Ix}}$$

$dout @ (K^T)^T \to K$

• Embedding affects entire Ith row of output by ext @ each element jth row of matrix. (Induced for all J, so summed)

$$\frac{\partial L}{\partial Emb_{in}}$$

↘

$$\sum_x \frac{\partial L}{\partial out_{Ix}} \cdot Mat_{Jx} = \frac{\partial L}{\partial V} @ V_x^T + \frac{\partial L}{\partial K} @ K_x^T + \frac{\partial L}{\partial Q} @ Q_x^T$$

↑ Q, K, V