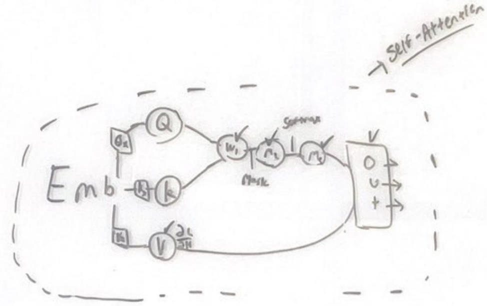
Deriving Self-Attention Head (w_1, m_1, m_2, V)

- Receive: $\frac{\partial L}{\partial out}$ → shape B, T, C

- Want: $\frac{\partial L}{\partial K}, \frac{\partial L}{\partial Q}, \frac{\partial L}{\partial V}$, → output $\frac{\partial L}{\partial \text{Prev (Emb)}}$

$\frac{\partial L}{\partial K}$ → (B, T, C) @ (C, H_size) (In ↑ above (B, T, C); key ↑ above (C, H_size))

- ∂ceH... affects each Batch-matrix



$m @ V$

$B\left[T[\diagdown^{T}] \ldots \right] @ T[\diagdown^{H}] \rightarrow out^{B}\left(T[\ ^{H}] \ldots \right.$

$\frac{\partial L}{\partial H}$ → • Each El_{IJ} affects all Batches (sum)

- Each El_{IJ} affects – Ith col of m – Jth col of out!

$\frac{\partial L}{\partial EL_{IJ}} \rightarrow \sum_{B} \sum_{x} \frac{\partial L}{\partial out_{xJ}} \cdot m_{xI}$

↓

$\boxed{m^T @ \frac{\partial L}{\partial out}}$.sum(0)

$\frac{\partial L}{\partial m}$ → • Each El_{IJ} affects corr. B', i: row$_I$ of output by row$_J$ of V → $\sum_{x} \frac{\partial L}{\partial out_{Ix}} \cdot H_{Jx}$

$\frac{\partial L}{\partial m_2} = \boxed{\frac{\partial L}{\partial out} @ V^T}$

↓

$\frac{\partial L}{\partial m_{1i}} \rightarrow y = \frac{e_i^x}{\sum_J e_J^x} \rightarrow -\frac{e^{2x}}{(\sum_J e_J^x)^2} + \frac{e_i^x}{\sum_J e_J^x}$

$-\frac{e_n^x \cdot e_i^x}{(\sum e_J^x)^2}$

$-\frac{e_n^x}{\sum_J e_J^x} \cdot \frac{e_i^x}{\sum_J e_J^x}$

Itself: $softmax_i - softmax_i^2$ $\cdot \frac{\partial L}{\partial m_2}$

Others: $-softmax_i \sum_n softmax_n$ $\cdot \frac{\partial L}{\partial m_2}$

- Sum collumns, multi by $-softm_i$ (El) → mul by $\frac{\partial L}{\partial m_2}$
- for itself, add $softmax_i$!

$\frac{\partial L}{\partial w}$ → just "stop" gradients to be set to zero!