$$S = QK^T$$

$$P = \text{softmax}(S)$$

$$O = PV$$

$$\begin{bmatrix} [P_{11} & P_{12} & P_{13}] \\ & & \\ \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} & V_{13} & V_{14} \\ V_{21} & V_{22} & V_{23} & V_{24} \\ V_{31} & V_{32} & V_{33} & V_{34} \end{bmatrix}$$

1st row

$$= \begin{bmatrix} (P_{11}V_{11} + P_{12}V_{21} + P_{13}V_{31}) & (P_{11}V_{12} + P_{12}V_{22} + P_{13}V_{32}) & (P_{11}V_{13} + P_{12}V_{23} + P_{13}V_{23}) & (P_{11}V_{14} + P_{12}V_{24} + P_{13}V_{34}) \\ & & & \end{bmatrix}$$

$$O_1 = P_{11}[V_{1:}] + P_{12}[V_{2:}] + P_{13}[V_{3:}]$$

$$\boxed{O_i = \sum_j P_{ij}(V_j)}$$

$$Y = XW$$

$$\frac{d\phi}{dX} = dX = \frac{d\phi}{dY} W^T$$

$$\frac{d\phi}{dW} = dW = X^T \frac{d\phi}{dY}$$

When $O = PV$

$$\boxed{dP = dO \cdot V^T}$$

$$\boxed{dV = P^T dO}$$

$$dV_i = \sum_j (P^T)_{ij} \cdot dO_j$$

$$\boxed{dV_i = \sum_j P_{ji}\, dO_j}$$

$$\boxed{dP_i = \sum_j dO_{ij} (V^T)_j}$$

$$P_{i:} = \text{softmax}(S_{i:})$$

For a vector x, y, s.t. $y = \text{softmax}(x)$

Jacobian $= \boxed{\text{diag}(y) - y \cdot y^T}$ $\longrightarrow$ Symmetric

$$dx = dS_{i:} = \left(\text{diag}(P_{i:}) - P_{i:} P_{i:}^T\right)$$

$$\frac{d\phi}{dS} = \frac{d\phi}{dP} \cdot \frac{dP}{dS} \longrightarrow \text{Row convention}$$

$\Rightarrow$ This is in column convention

$$\Rightarrow \left(\frac{d\phi}{dS}\right)^T = \underbrace{\frac{dP}{dS}}_{\text{Symmetric}} \cdot \underbrace{\left(\frac{d\phi}{dP}\right)^T}_{= dP_{i:}}$$

$$dS_{i:} = \left(\mathrm{diag}(P_{i:}) - P_{i:}P_{i:}^T\right) dP_{i:}$$

$$= \left(P_{i:} \odot dP_{i:}\right) - \underbrace{\left(P_{i:}^T dP_{i:}\right) P_{i:}}_{\text{Strange but equivalent}}$$

Define

$$D_i = P_{i:}^T \, dP_{i:}$$

$$D_i = \sum_j \left(P_{i:}^T\right)_{ij} \left(dP_{i:}\right)_j$$

$$= \sum_j \frac{e^{q_i^T k_j}}{L_i} \cdot \underbrace{dO_i^T \cdot V_j}_{1\times 1} = dO_i^T \sum_j \frac{e^{q_i^T k_j}}{L_i} V_j$$

$$= dO_i^T \cdot O_i$$

$$\Rightarrow \boxed{D_i = dO_i^T O_i}$$

$$\boxed{dS_{i:} = P_{i:} \odot dP_{i:} - D_i P_{i:}}$$

$$dS_{ij} = \left(P_{ij} \times dP_{ij}\right) - D_i P_{ij}$$

$$\Rightarrow \boxed{dS_{ij} = P_{ij}\left(dP_{ij} - D_i\right)}$$

Now, $S_{ij} = q_i^T k_j \Rightarrow S = QK^T$

$$\boxed{\begin{aligned} dQ &= dS \cdot K \\ dK &= dS^T \cdot Q \end{aligned}}$$

$$\boxed{\begin{aligned} dq_i = dS_{i:} K &= \sum_j dS_{ij} k_j \\ &= \sum_j P_{ij}\left(dP_{ij} - D_i\right) k_j \\ &= \sum_j \frac{e^{q_i^T k_j}}{L_i}\left(do_i^T v_j - D_i\right) k_j \end{aligned}}$$

Similarly

$$\boxed{\begin{aligned} dk_j = \sum_i \underbrace{dS_{ij}}_{\text{After Transpose}} q_i &= \sum_i P_{ij}\left(dP_{ij} - D_i\right) q_i \\ &= \sum_i \frac{e^{q_i^T k_j}}{L_i}\left(do_i^T v_j - D_i\right) q_i \end{aligned}}$$