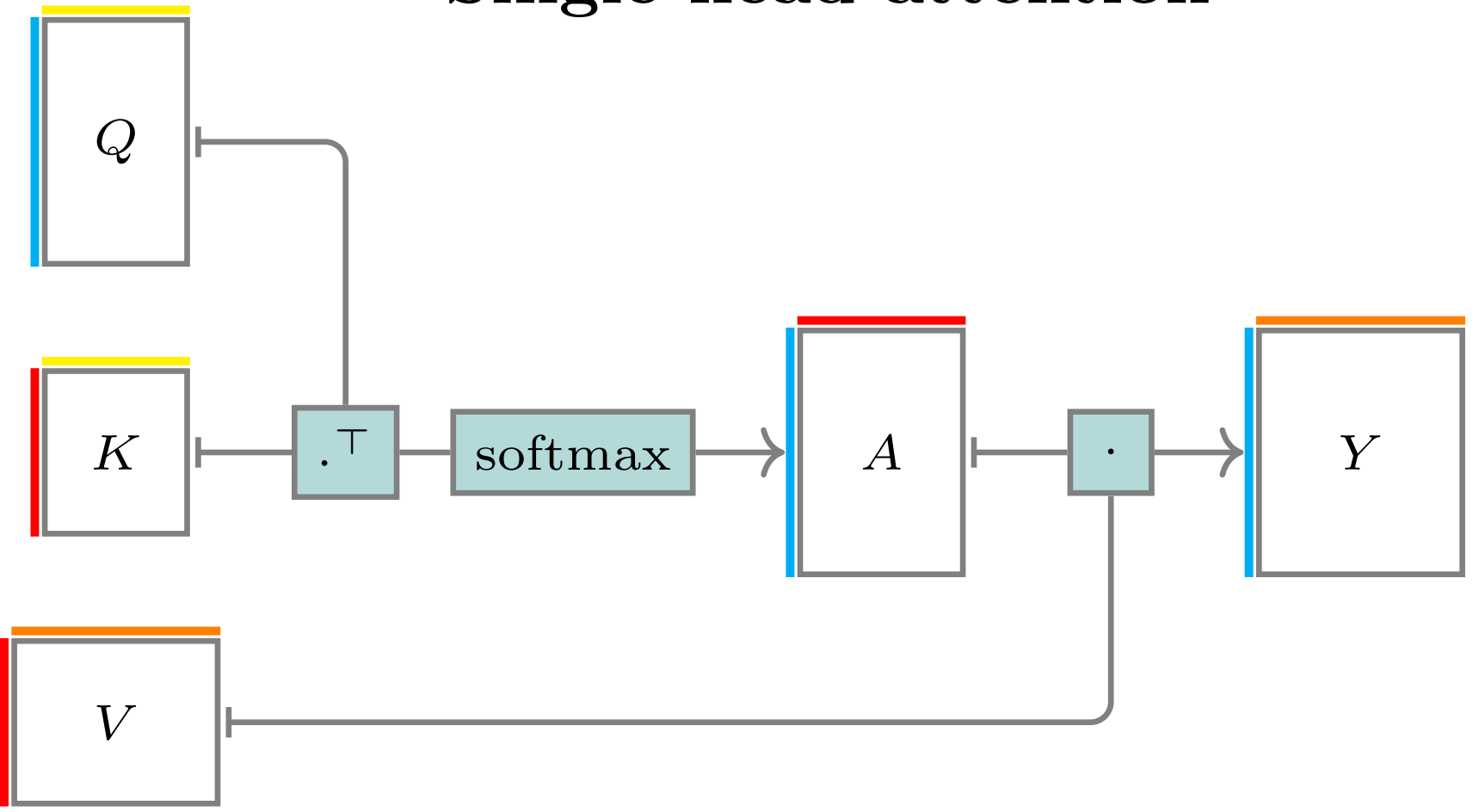
Single-head attention



$$\text{Attention}(Q, K, V) = \text{softmax}_{\text{row}}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$