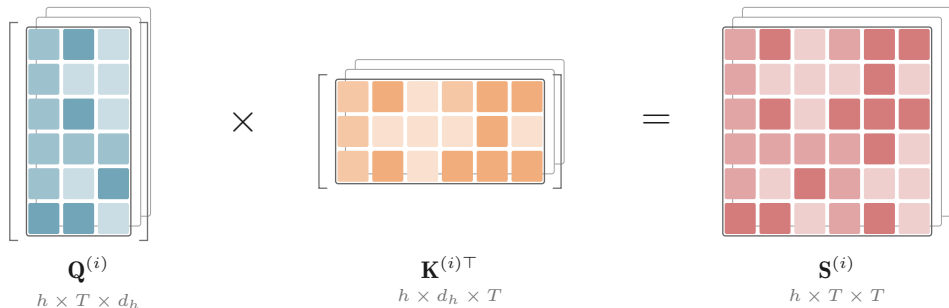


$$S^{(i)} = Q^{(i)} K^{(i)\top}$$

每头打分：沿 d_h 收缩



轴	T 时间； d_h 头维； h 头数画成 stack 深度
对象	Q/K 是 value； S 是 score，不是 mask
机制	沿 d_h 收缩； $K^\top$ 换面，收缩边等长