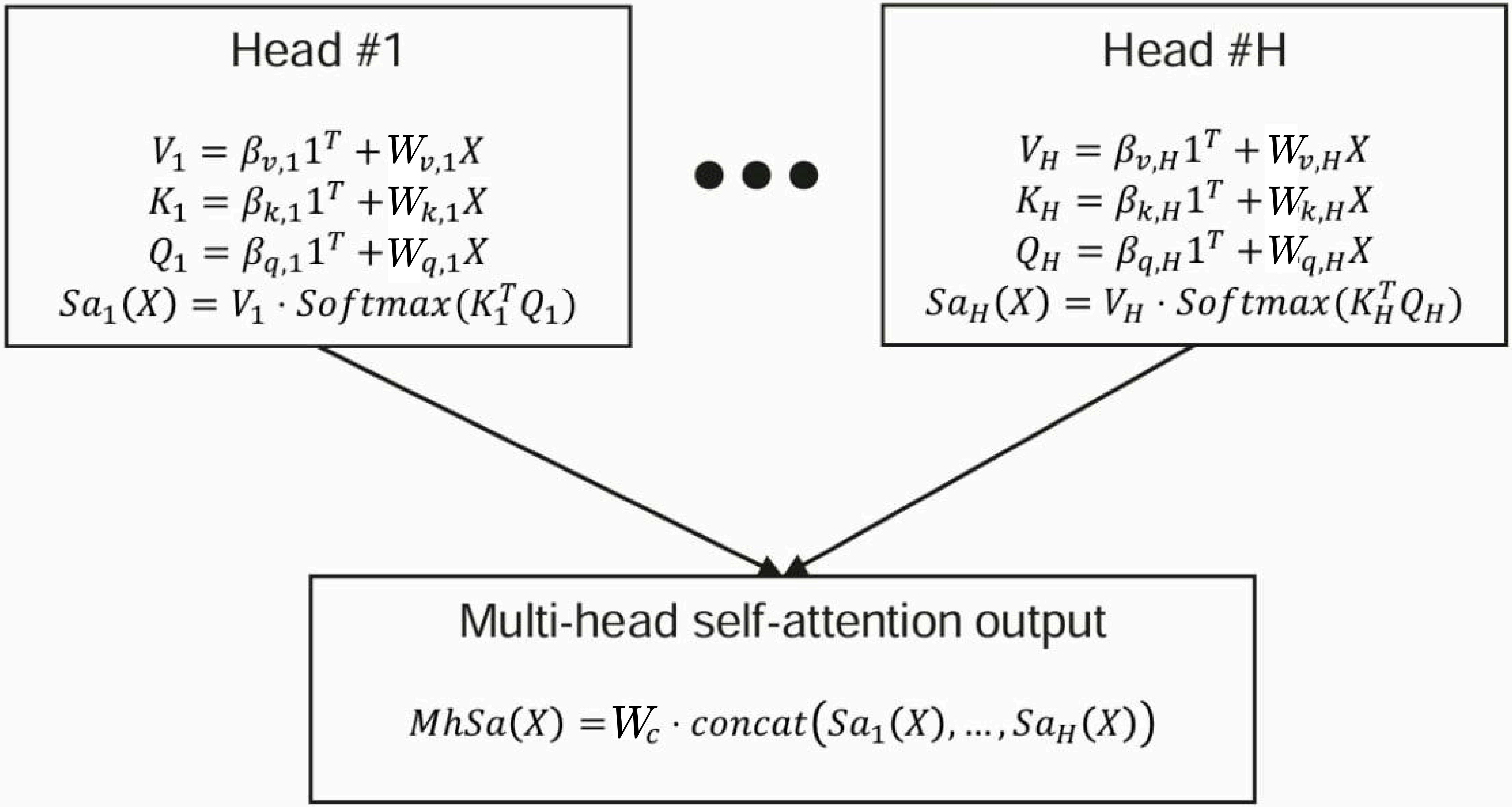

Head #1

$$V_1 = \beta_{v,1} 1^T + W_{v,1} X$$
$$K_1 = \beta_{k,1} 1^T + W_{k,1} X$$
$$Q_1 = \beta_{q,1} 1^T + W_{q,1} X$$
$$Sa_1(X) = V_1 \cdot Softmax(K_1^T Q_1)$$

●●●

Head #H

$$V_H = \beta_{v,H} 1^T + W_{v,H} X$$
$$K_H = \beta_{k,H} 1^T + W_{k,H} X$$
$$Q_H = \beta_{q,H} 1^T + W_{q,H} X$$
$$Sa_H(X) = V_H \cdot Softmax(K_H^T Q_H)$$

Multi-head self-attention output

$$MhSa(X) = W_c \cdot concat\big(Sa_1(X), \ldots, Sa_H(X)\big)$$