



Figure 1: Multi-head Attention.

$$Q^{[D],[L]} = \text{prior} \quad (1a)$$

$$K^{[D],[L]} = \text{prior} \quad (1b)$$

$$V^{[D],[L]} = \text{prior} \quad (1c)$$

$$Q_0^{[d],[L]} = \text{linear}(Q^{[D],[L]}) \quad (1d)$$

$$Q_1^{[d],[L]} = \text{linear}(Q^{[D],[L]}) \quad (1e)$$

$$Q_2^{[d],[L]} = \text{linear}(Q^{[D],[L]}) \quad (1f)$$

$$K_0^{[d],[L]} = \text{linear}(K^{[D],[L]}) \quad (1g)$$

$$K_1^{[d],[L]} = \text{linear}(K^{[D],[L]}) \quad (1h)$$

$$K_2^{[d],[L]} = \text{linear}(K^{[D],[L]}) \quad (1i)$$

$$V_0^{[d],[L]} = \text{linear}(V^{[D],[L]}) \quad (1j)$$

$$V_1^{[d],[L]} = \text{linear}(V^{[D],[L]}) \quad (1k)$$

$$V_2^{[d],[L]} = \text{linear}(V^{[D],[L]}) \quad (1l)$$

$$A_0^{[d],[L]} = \text{scaled_dot_prod_att}(Q_0^{[d],[L]}, K_0^{[d],[L]}, V_0^{[d],[L]}) \quad (1m)$$

$$A_1^{[d],[L]} = \text{scaled_dot_prod_att}(Q_1^{[d],[L]}, K_1^{[d],[L]}, V_1^{[d],[L]}) \quad (1n)$$

$$A_2^{[d],[L]} = \text{scaled_dot_prod_att}(Q_2^{[d],[L]}, K_2^{[d],[L]}, V_2^{[d],[L]}) \quad (1o)$$

$$A^{[D],[L]} = [A_0^{[d],[L]} | A_1^{[d],[L]} | A_2^{[d],[L]}] \quad (1p)$$

$$F^{[D],[L]} = W_{\underline{c}}^{[D][D]} A^{[D],[L]} \quad (1q)$$