



Figure 1: Multi-head Attention.

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

$$A_0^{[d],[\ell]} = \text{scaled\_dot\_prod\_att}(Q_0^{[d],[\ell]}, K_0^{[d],[\ell]}, V_0^{[d],[\ell]}) \quad (1b)$$

$$A_1^{[d],[\ell]} = \text{scaled\_dot\_prod\_att}(Q_1^{[d],[\ell]}, K_1^{[d],[\ell]}, V_1^{[d],[\ell]}) \quad (1c)$$

$$A_2^{[d],[\ell]} = \text{scaled\_dot\_prod\_att}(Q_2^{[d],[\ell]}, K_2^{[d],[\ell]}, V_2^{[d],[\ell]}) \quad (1d)$$

$$K^{[D],[\ell]} = prior \quad (1e)$$

$$K_0^{[d],[\ell]} = \text{linear}(K^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1f)$$

$$K_1^{[d],[\ell]} = \text{linear}(K^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1g)$$

$$K_2^{[d],[\ell]} = \text{linear}(K^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1h)$$

$$O^{[D],[\ell]} = W_{\underline{2}}^{[D],[D]} A^{[D],[\ell]} \quad (1i)$$

$$Q^{[D],[\ell]} = prior \quad (1j)$$

$$Q_0^{[d],[\ell]} = \text{linear}(Q^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1k)$$

$$Q_1^{[d],[\ell]} = \text{linear}(Q^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1l)$$

$$Q_2^{[d],[\ell]} = \text{linear}(Q^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1m)$$

$$V^{[D],[\ell]} = prior \quad (1n)$$

$$V_0^{[d],[\ell]} = \text{linear}(V^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1o)$$

$$V_1^{[d],[\ell]} = \text{linear}(V^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1p)$$

$$V_2^{[d],[\ell]} = \text{linear}(V^{[D],[\ell]}) \text{ (split, then project a component)} \quad (1q)$$