



Figure 1: Multi-head Attention with 2 heads. Note that the input \underline{e} and output \underline{Q} have the same shape.

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

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

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

$$K^{[D],[\ell]} = W_{\underline{k}}^{[D],[d]} \underline{e}^{[d],[\ell]} \quad (1d)$$

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

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

$$O^{[d],[\ell]} = W_{\underline{a}}^{[d],[D]} A^{[D],[\ell]} \quad (1g)$$

$$Q^{[D],[\ell]} = W_{\underline{q}}^{[D],[d]} e^{[d],[\ell]} \quad (1h)$$

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

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

$$V^{[D],[\ell]} = W_{\underline{v}}^{[D],[d]} e^{[d],[\ell]} \quad (1k)$$

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

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

$$e^{[d],[\ell]} = \text{prior} \quad (1n)$$