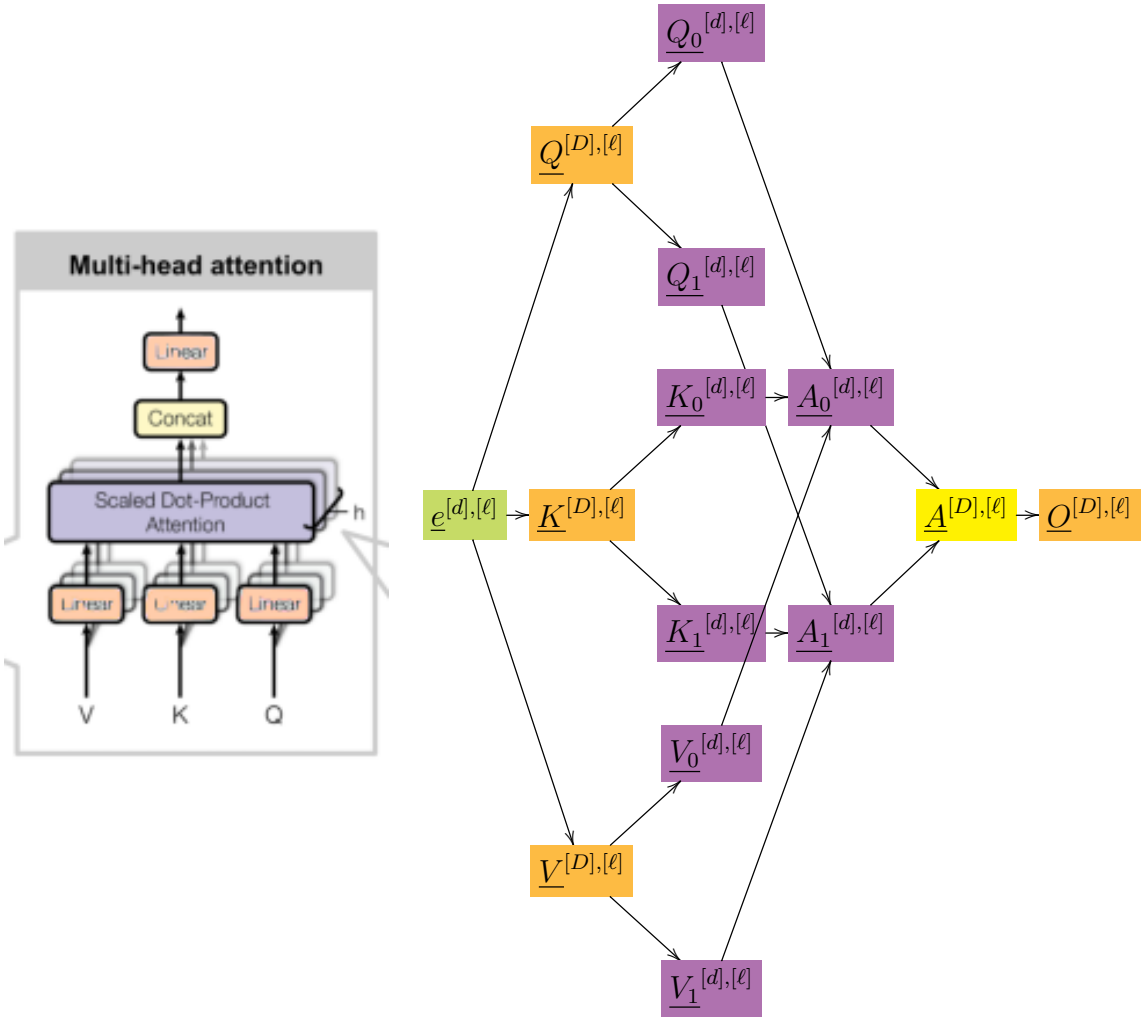

Figure 1: Multi-head Attention with 2 heads.

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

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

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

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

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

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

$$O^{[D],[\ell]} = W_{\underline{o}}^{[D],[D]} A^{[D],[\ell]} \tag{1g}$$

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

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

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

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

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

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

$$e^{[d],[\ell]} = \;\; \text{prior} \tag{1n}$$