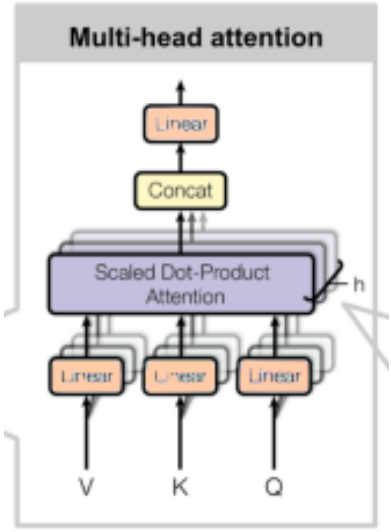

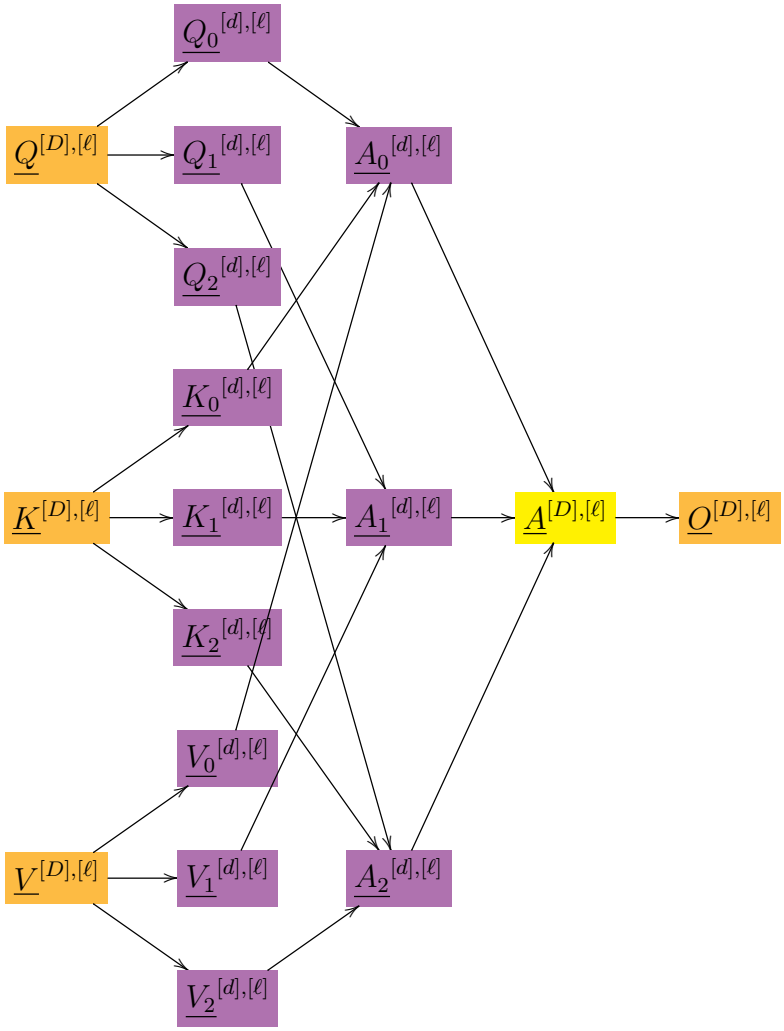

Figure 1: Multi-head Attention.

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

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

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

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

$$K^{[D],[\ell]} = prior \tag{1e}$$

$$K_0^{[d],[\ell]} = \text{linear}(K^{[D],[\ell]}) \tag{1f}$$

$$K_1^{[d],[\ell]} = \text{linear}(K^{[D],[\ell]}) \tag{1g}$$

$$K_2^{[d],[\ell]} = \text{linear}(K^{[D],[\ell]}) \tag{1h}$$

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

$$Q^{[D],[\ell]} = prior \tag{1j}$$

$$Q_0^{[d],[\ell]} = \text{linear}(Q^{[D],[\ell]}) \tag{1k}$$

$$Q_1^{[d],[\ell]} = \text{linear}(Q^{[D],[\ell]}) \tag{1l}$$

$$Q_2^{[d],[\ell]} = \text{linear}(Q^{[D],[\ell]}) \tag{1m}$$

$$V^{[D],[\ell]} = prior \tag{1n}$$

$$V_0^{[d],[\ell]} = \text{linear}(V^{[D],[\ell]}) \tag{1o}$$

$$V_1^{[d],[\ell]} = \text{linear}(V^{[D],[\ell]}) \tag{1p}$$

$$V_2^{[d],[\ell]} = \text{linear}(V^{[D],[\ell]}) \tag{1q}$$