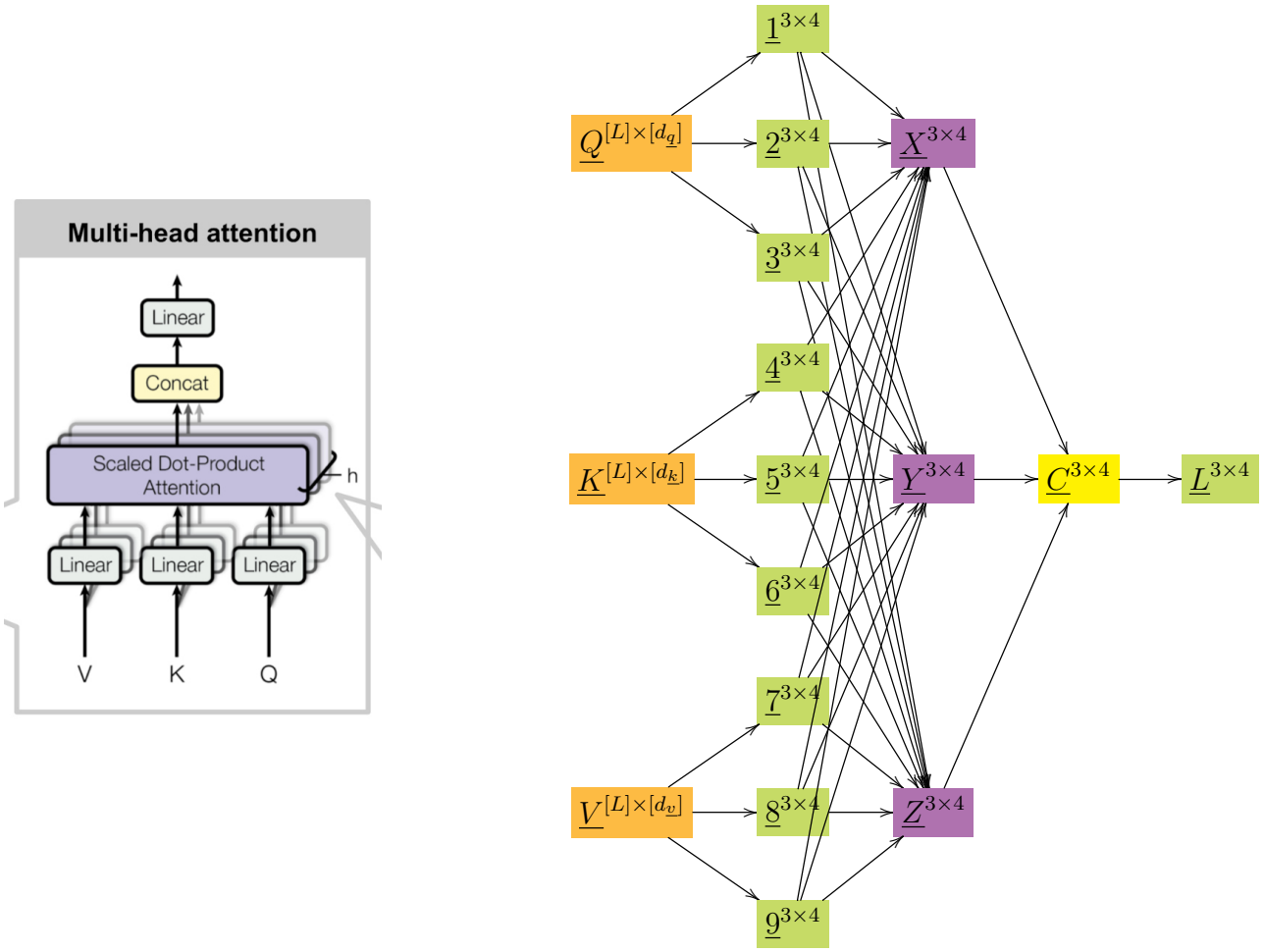

Figure 1: Multi-head Attention.

$$Q^{[L]\times[d_{\underline{q}}]} = prior \tag{1a}$$

$$K^{[L]\times[d_{\underline{k}}]} = prior \tag{1b}$$

$$V^{[L]\times[d_{\underline{v}}]} = prior \tag{1c}$$

$$1^{3\times4} = \text{linear}(Q^{[L]\times[d_{\underline{q}}]}) \tag{1d}$$

$$2^{3\times4} = \text{linear}(Q^{[L]\times[d_{\underline{q}}]}) \tag{1e}$$

$$3^{3\times4} = \text{linear}(Q^{[L]\times[d_{\underline{q}}]}) \tag{1f}$$

$$4^{3\times4} = \text{linear}(K^{[L]\times[d_{\underline{k}}]}) \tag{1g}$$

$$5^{3\times4} = \text{linear}(K^{[L]\times[d_{\underline{k}}]}) \tag{1h}$$

$$6^{3\times4} = \text{linear}(K^{[L]\times[d_{\underline{k}}]}) \tag{1i}$$

$$7^{3\times4} = \text{linear}(V^{[L]\times[d_{\underline{v}}]}) \tag{1j}$$

$$8^{3\times4} = \text{linear}(V^{[L]\times[d_{\underline{v}}]}) \tag{1k}$$

$$9^{3\times4} = \text{linear}(V^{[L]\times[d_{\underline{v}}]}) \tag{1l}$$

$$X^{3\times4} = \text{scaled\_dot\_prod\_att}(1^{3\times4}, 2^{3\times4}, 3^{3\times4}, 4^{3\times4}, 5^{3\times4}, 6^{3\times4}, 7^{3\times4}, 8^{3\times4}, 9^{3\times4}) \tag{1m}$$

$$Y^{3\times4} = \text{scaled\_dot\_prod\_att}(1^{3\times4}, 2^{3\times4}, 3^{3\times4}, 4^{3\times4}, 5^{3\times4}, 6^{3\times4}, 7^{3\times4}, 8^{3\times4}, 9^{3\times4}) \tag{1n}$$

$$Z^{3\times4} = \text{scaled\_dot\_prod\_att}(1^{3\times4}, 2^{3\times4}, 3^{3\times4}, 4^{3\times4}, 5^{3\times4}, 6^{3\times4}, 7^{3\times4}, 8^{3\times4}, 9^{3\times4}) \tag{1o}$$

$$C^{3\times4} = [X^{3\times4}|Y^{3\times4}|Z^{3\times4}] \tag{1p}$$

$$L^{3\times4} = C^{3\times4}W_{\underline{o}}^{[d_{\underline{v}}]\times[d]} \tag{1q}$$