



Figure 1: Multi-head Attention.

$$\underline{Q}^{3 \times 4} = ) \quad (1a)$$

$$\underline{K}^{3 \times 4} = ) \quad (1b)$$

$$\underline{V}^{3 \times 4} = ) \quad (1c)$$

$$\underline{1}^{3 \times 4} = \text{linear}(\underline{Q}^{3 \times 4}) \quad (1d)$$

$$\underline{2}^{3 \times 4} = \text{linear}(\underline{Q}^{3 \times 4}) \quad (1e)$$

$$\underline{3}^{3 \times 4} = \text{linear}(\underline{Q}^{3 \times 4}) \quad (1f)$$

$$\underline{4}^{3 \times 4} = \text{linear}(\underline{K}^{3 \times 4}) \quad (1g)$$

$$\underline{5}^{3 \times 4} = \text{linear}(\underline{K}^{3 \times 4}) \quad (1h)$$

$$6^{3 \times 4} = \text{linear}(K^{3 \times 4}) \quad (1i)$$

$$7^{3 \times 4} = \text{linear}(V^{3 \times 4}) \quad (1j)$$

$$8^{3 \times 4} = \text{linear}(V^{3 \times 4}) \quad (1k)$$

$$9^{3 \times 4} = \text{linear}(V^{3 \times 4}) \quad (1l)$$

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

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

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

$$C^{3 \times 4} = \text{concat}(X^{3 \times 4}, Y^{3 \times 4}, Z^{3 \times 4}) \quad (1p)$$

$$L^{3 \times 4} = \text{concat}(C^{3 \times 4}) \quad (1q)$$