



Figure 1: Multi-head Attention.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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