



Figure 1: Encoder.

$$F^{[\Lambda],[D],[\ell]} = \text{feed_forward_nn}(N^{[\Lambda],[D],[\ell]}) \quad (1a)$$

$$K^{[\Lambda],[D],[\ell]} = W_{\underline{k}}^{[D],[d]} e^{[\Lambda],[d],[\ell]} \quad (1b)$$

$$N^{[\Lambda],[D],[\ell]} = \text{normalize}(e^{[\Lambda],[d],[\ell]} + O^{[\Lambda],[D],[\ell]}) \quad (1c)$$

$$O^{[\Lambda],[D],[\ell]} = \text{multi_headed_attention}(Q^{[\Lambda],[D],[\ell]}, K^{[\Lambda],[D],[\ell]}, V^{[\Lambda],[D],[\ell]}) \quad (1d)$$

$$Q^{[\Lambda],[D],[\ell]} = W_{\underline{q}}^{[D],[d]} e^{[\Lambda],[d],[\ell]} \quad (1e)$$

$$V^{[\Lambda],[D],[\ell]} = W_{\underline{v}}^{[D],[d]} e^{[\Lambda],[d],[\ell]} \quad (1f)$$

$$e^{[\Lambda],[d],[\ell]} = E^{[\Lambda],[d],[L]} x^{[L],[\ell]} \quad (1g)$$

$$k^{[\Lambda],[D],[\ell]} = n^{[\Lambda],[D],[\ell]} \quad (1h)$$

$$n^{[\Lambda],[D],[\ell]} = \text{normalize}(N^{[\Lambda],[D],[\ell]} + F^{[\Lambda],[D],[\ell]}) \quad (1i)$$

$$q^{[\Lambda],[D],[\ell]} = n^{[\Lambda],[D],[\ell]} \quad (1j)$$

$$x^{[L],[\ell]} = \text{prior} \quad (1k)$$