



Figure 1: Encoder.

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

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

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

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

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

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

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

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

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

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

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