Figure 1: Encoder.

$$F^{[\Lambda],[D],[\ell]} = \text{feed\_forward\_nn}(N^{[\Lambda],[D],[\ell]}) \tag{1a}$$

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

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

1

$$O^{[\Lambda],[D],[\ell]} = \text{multi\_headed\_attention}(Q^{[\Lambda],[D],[\ell]}, K^{[\Lambda],[D],[\ell]}, V^{[\Lambda],[D],[\ell]}) \tag{1d}$$

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

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

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

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

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

$$v^{[\Lambda],[D],[\ell]} = n^{[\Lambda],[D],[\ell]} \tag{1j}$$

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