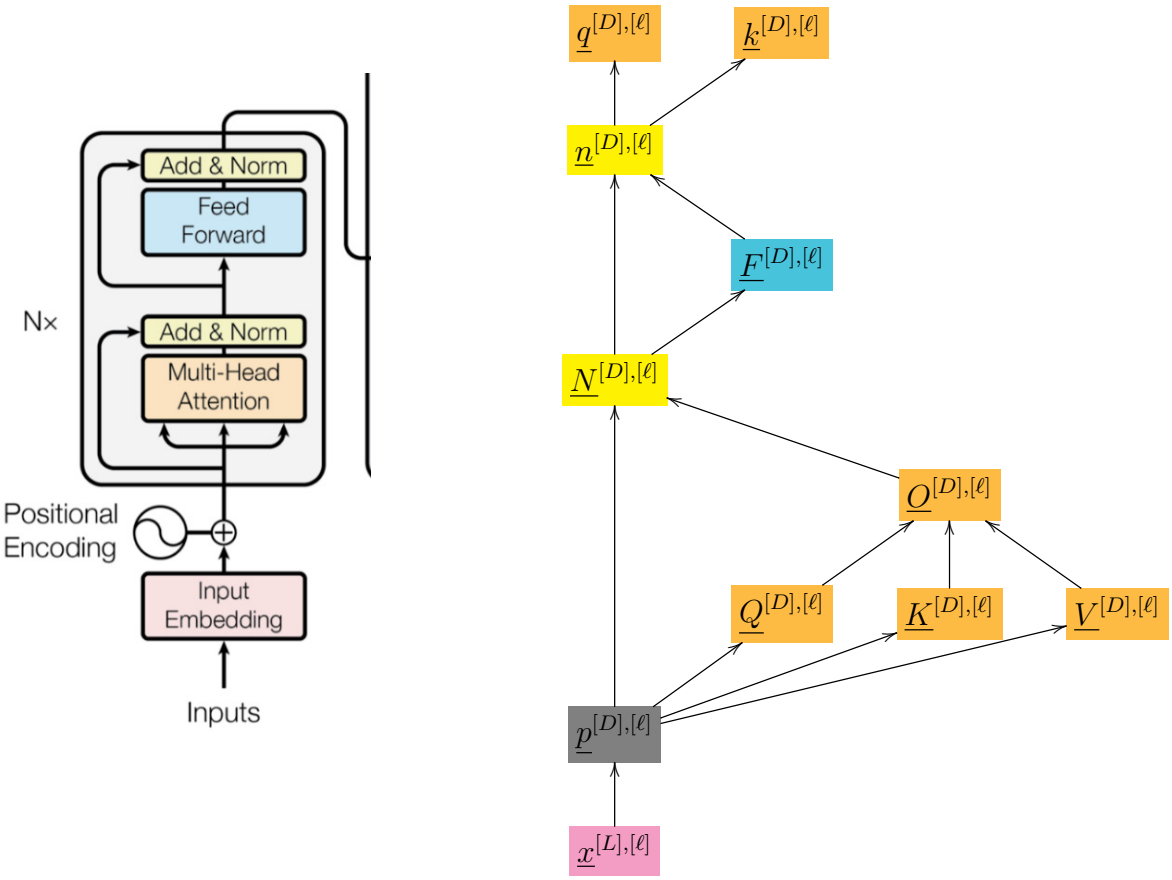

Figure 1: Encoder.

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

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

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

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

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

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

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

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

$$p^{[D],[\ell]} = M^{[\ell],[\ell]} x^{[L],[\ell]} \tag{1i}$$

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

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