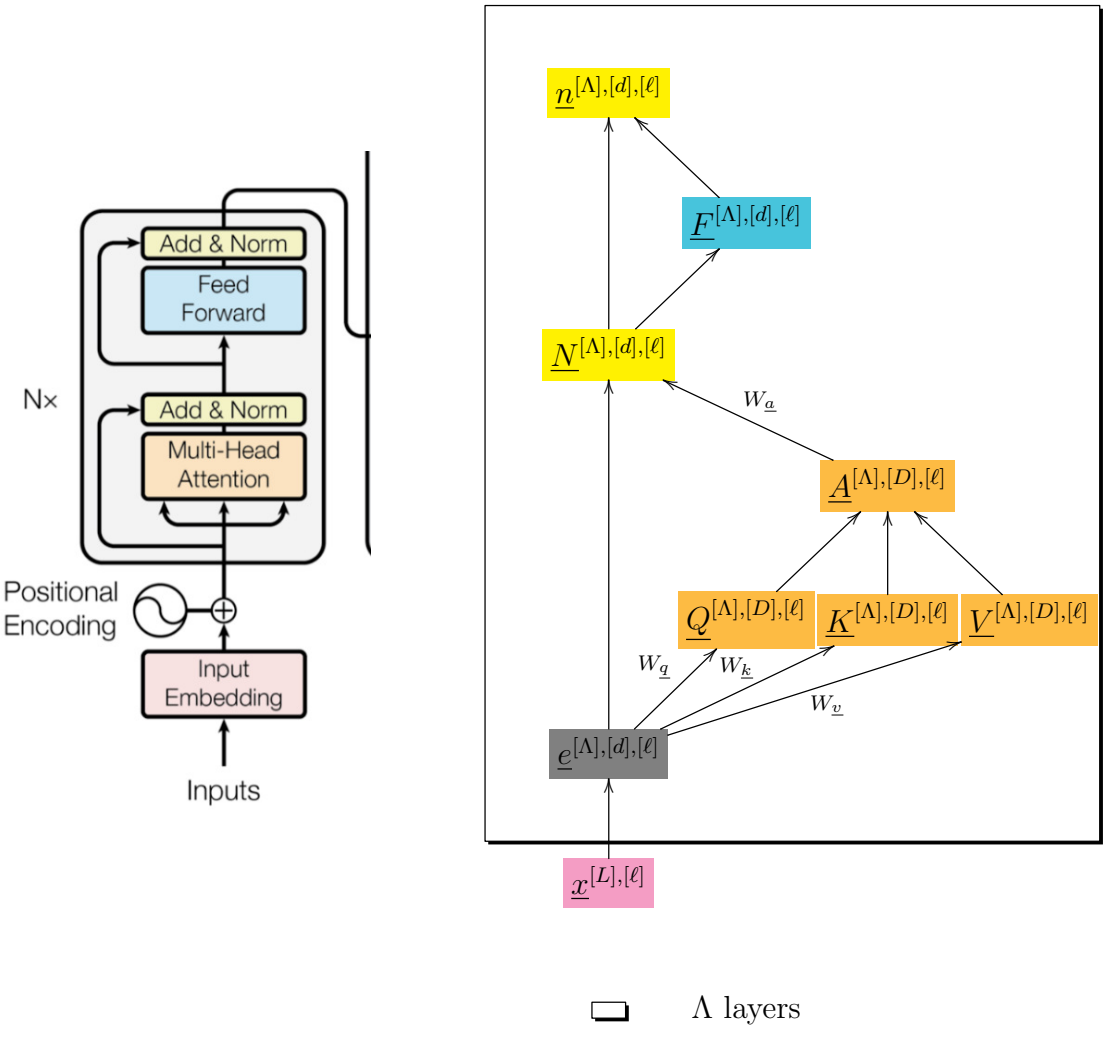

Figure 1: Encoder of Vanilla Transform Net.

$$A^{[\Lambda],[D],[\ell]} = \text{Attention}(Q^{[\Lambda],[D],[\ell]}, K^{[\Lambda],[D],[\ell]}, V^{[\Lambda],[D],[\ell]}) \tag{1a}$$

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

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

$$N^{[\Lambda],[d],[\ell]} = \text{normalize}(e^{[\Lambda],[d],[\ell]} + W_{\underline{a}}^{[d],[D]} A^{[\Lambda],[D],[\ell]}) \tag{1d}$$

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

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

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

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

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