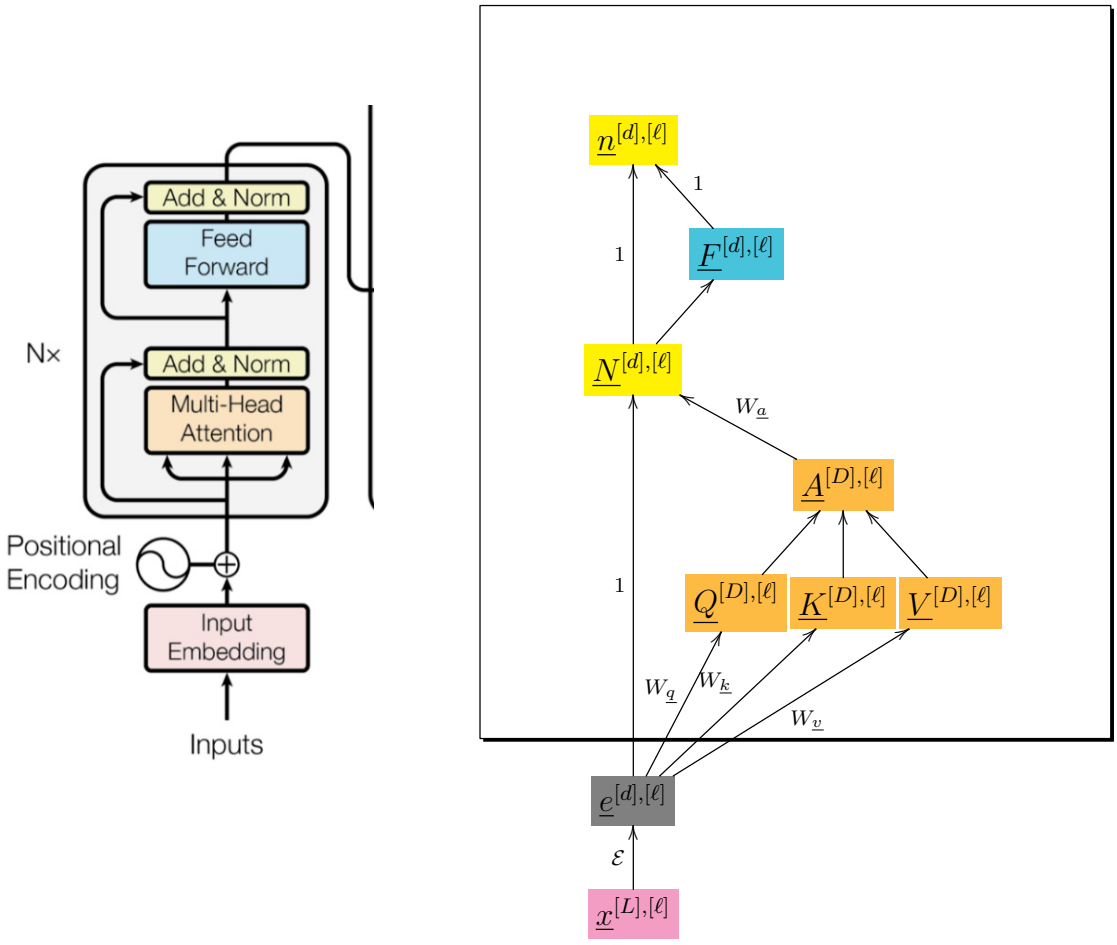

Figure 1: Encoder of Vanilla Transformer Net. Λ copies of the boxed part are connected in series.

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

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

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

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

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

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

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

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

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