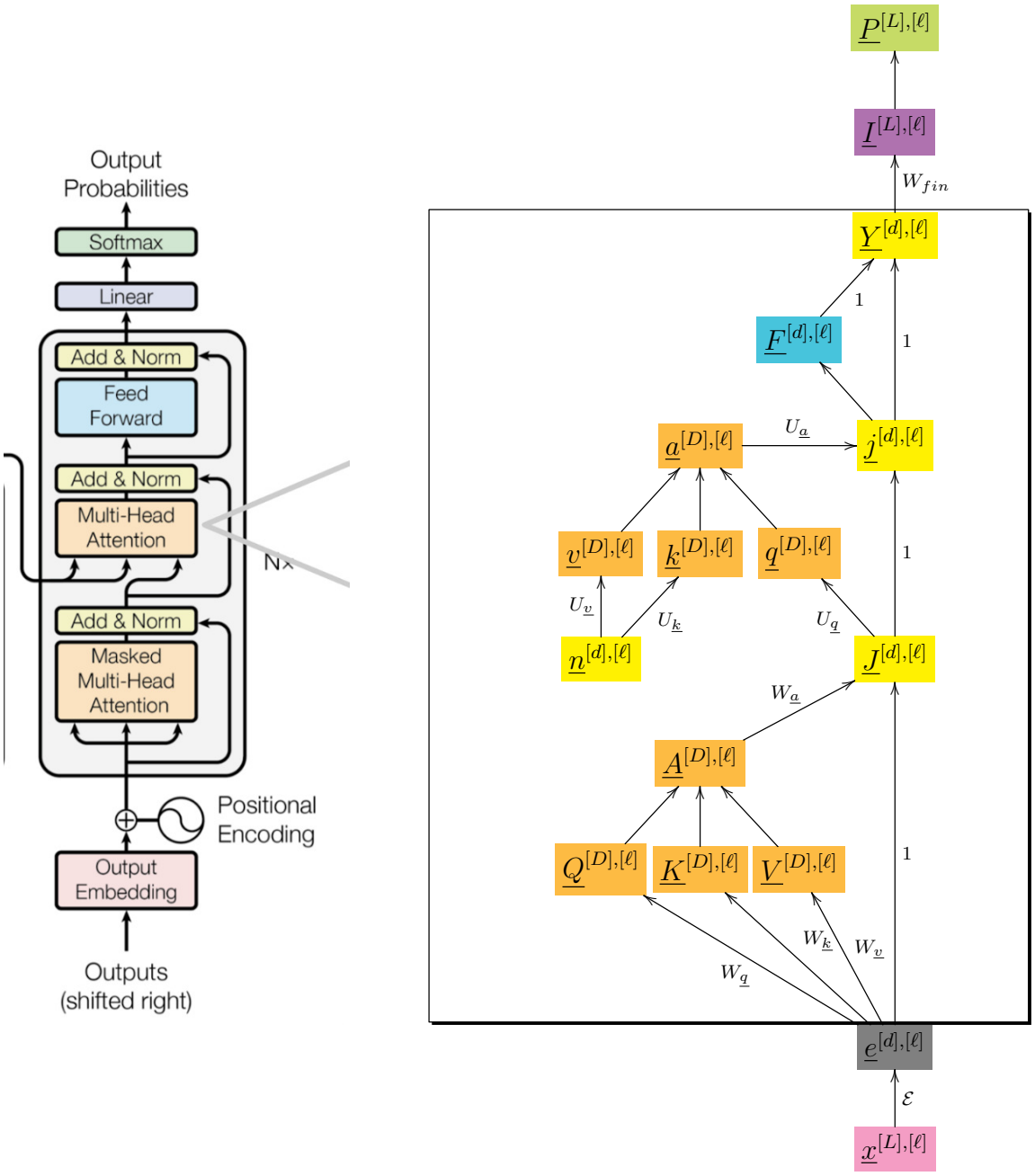


Figure 1: Decoder of Vanilla Transformer Net. $\Lambda$ copies of the boxed part are connected in series.

$$a^{[D],[\ell]} = \text{Attention}(v^{[D],[\ell]}, k^{[D],[\ell]}, q^{[D],[\ell]}) \tag{1a}$$

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

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

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

$$I^{[L],[\ell]} = W_{fin}^{[L],[d]} Y^{[d],[\ell]} \tag{1e}$$

$$j^{[d],[\ell]} = \text{normalize}(U_{\underline{a}}^{[d],[D]} a^{[D],[\ell]} + J^{[d],[\ell]}) \tag{1f}$$

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

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

$$k^{[D],[\ell]} = U_{\underline{k}}^{[D],[d]} n^{[d],[\ell]} \tag{1i}$$

$$n^{[d],[\ell]} = \quad \text{Prior coming from Encoder.} \tag{1j}$$

$$P^{[L],[\ell]} = \text{softmax}(I^{[L],[\ell]}) \quad (\sum_{\alpha \in [\ell]} P^{[L],\alpha} = 1) \tag{1k}$$

$$q^{[D],[\ell]} = U_{\underline{q}}^{[D],[d]} J^{[d],[\ell]} \tag{1l}$$

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

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

$$v^{[D],[\ell]} = U_{\underline{v}}^{[D],[d]} n^{[d],[\ell]} \tag{1o}$$

$$x^{[L],[\ell]} = \text{ prior, right shifted output} \tag{1p}$$

$$Y^{[d],[\ell]} = \text{normalize}(F^{[d],[\ell]} + J^{[d],[\ell]}) \tag{1q}$$