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

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

$$I^{[L],[\ell]} = W^{[L],[\Lambda],[d]} Y^{[\Lambda],[d],[\ell]} \tag{1c}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

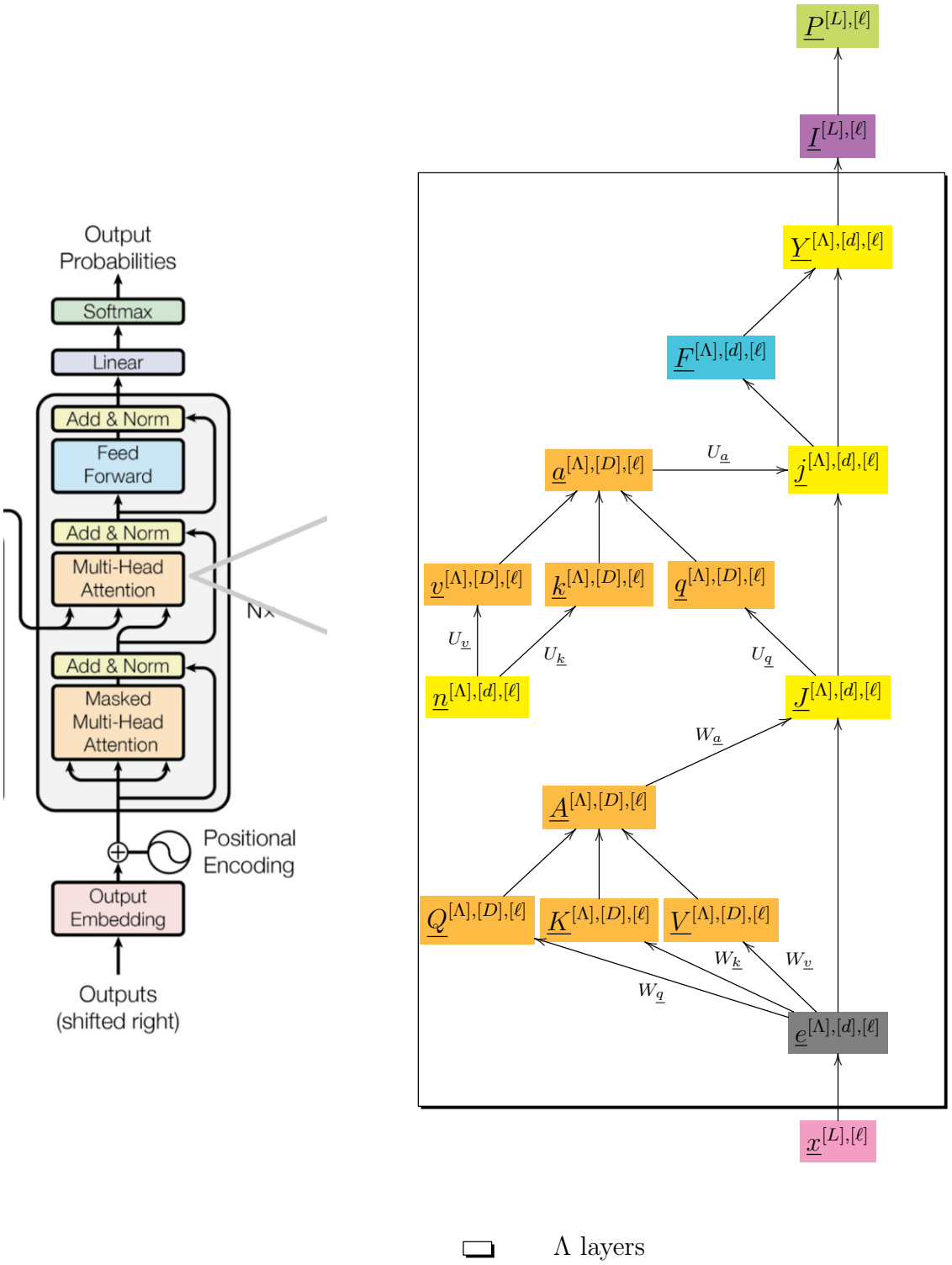

Figure 1: Decoder of Vanilla Transformer Net.