

$$F^{[\Lambda],[d],[\ell]} = \text{feed_forward_nn}(j^{[\Lambda],[d],[\ell]}) \quad (1a)$$

$$I^{[L],[\ell]} = W^{[L],[\Lambda],[d]} Y^{[\Lambda],[d],[\ell]} \quad (1b)$$

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

$$O^{[\Lambda],[d],[\ell]} = \text{multi_head_attention}(Q^{[\Lambda],[D],[\ell]}, K^{[\Lambda],[D],[\ell]}, V^{[\Lambda],[D],[\ell]}) \quad (1d)$$

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

$$Q^{[\Lambda],[D],[\ell]} = W_{\underline{q}}^{[D],[d]} e^{[\Lambda],[d],[\ell]} \quad (1f)$$

$$V^{[\Lambda],[D],[\ell]} = W_{\underline{v}}^{[D],[d]} e^{[\Lambda],[d],[\ell]} \quad (1g)$$

$$Y^{[\Lambda],[d],[\ell]} = \text{normalize}(F^{[\Lambda],[d],[\ell]} + a^{[\Lambda],[d],[\ell]}) \quad (1h)$$

$$a^{[\Lambda],[d],[\ell]} = \text{normalize}(O^{[\Lambda],[d],[\ell]} + e^{[\Lambda],[d],[\ell]}) \quad (1i)$$

$$e^{[\Lambda],[d],[\ell]} = E^{[\Lambda],[d],[L]} x^{[L],[\ell]} \quad (1j)$$

$$j^{[\Lambda],[d],[\ell]} = \text{normalize}(o^{[\Lambda],[d],[\ell]} + a^{[\Lambda],[d],[\ell]}) \quad (1k)$$

$$k^{[\Lambda],[D],[\ell]} = \text{prior, obtained from encoder.} \quad (1l)$$

$$o^{[\Lambda],[d],[\ell]} = \text{multi_head_attention}(v^{[\Lambda],[D],[\ell]}, k^{[\Lambda],[D],[\ell]}, q^{[\Lambda],[D],[\ell]}) \quad (1\text{m})$$

$$q^{[\Lambda],[D],[\ell]} = U_{\underline{q}}^{[D],[d]} a^{[\Lambda],[d],[\ell]} \quad (1\text{n})$$

$$v^{[\Lambda],[D],[\ell]} = \text{prior, obtained from encoder.} \quad (1\text{o})$$

$$x^{[L],[\ell]} = \text{prior} \quad (1\text{p})$$

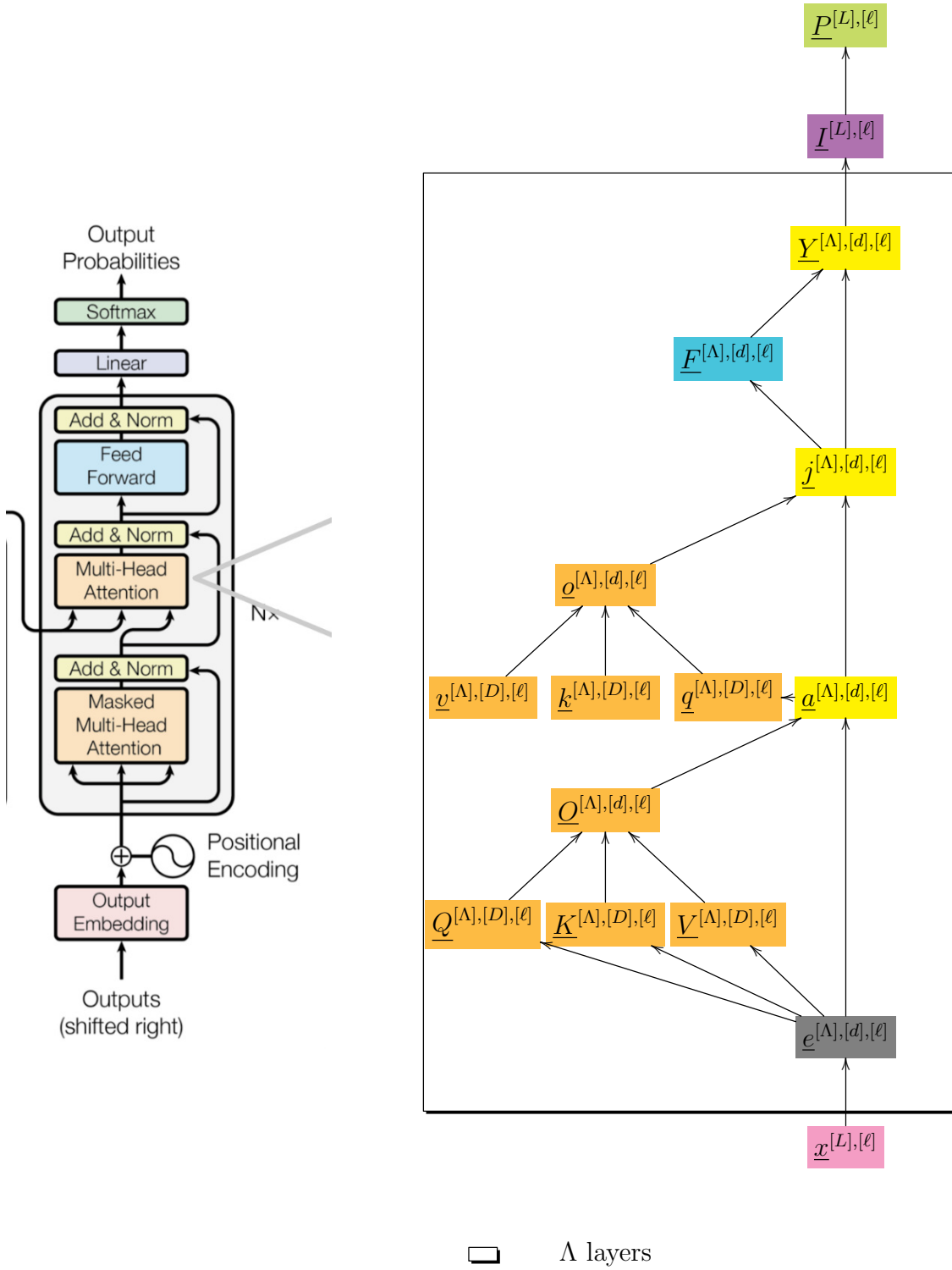


Figure 1: Decoder.