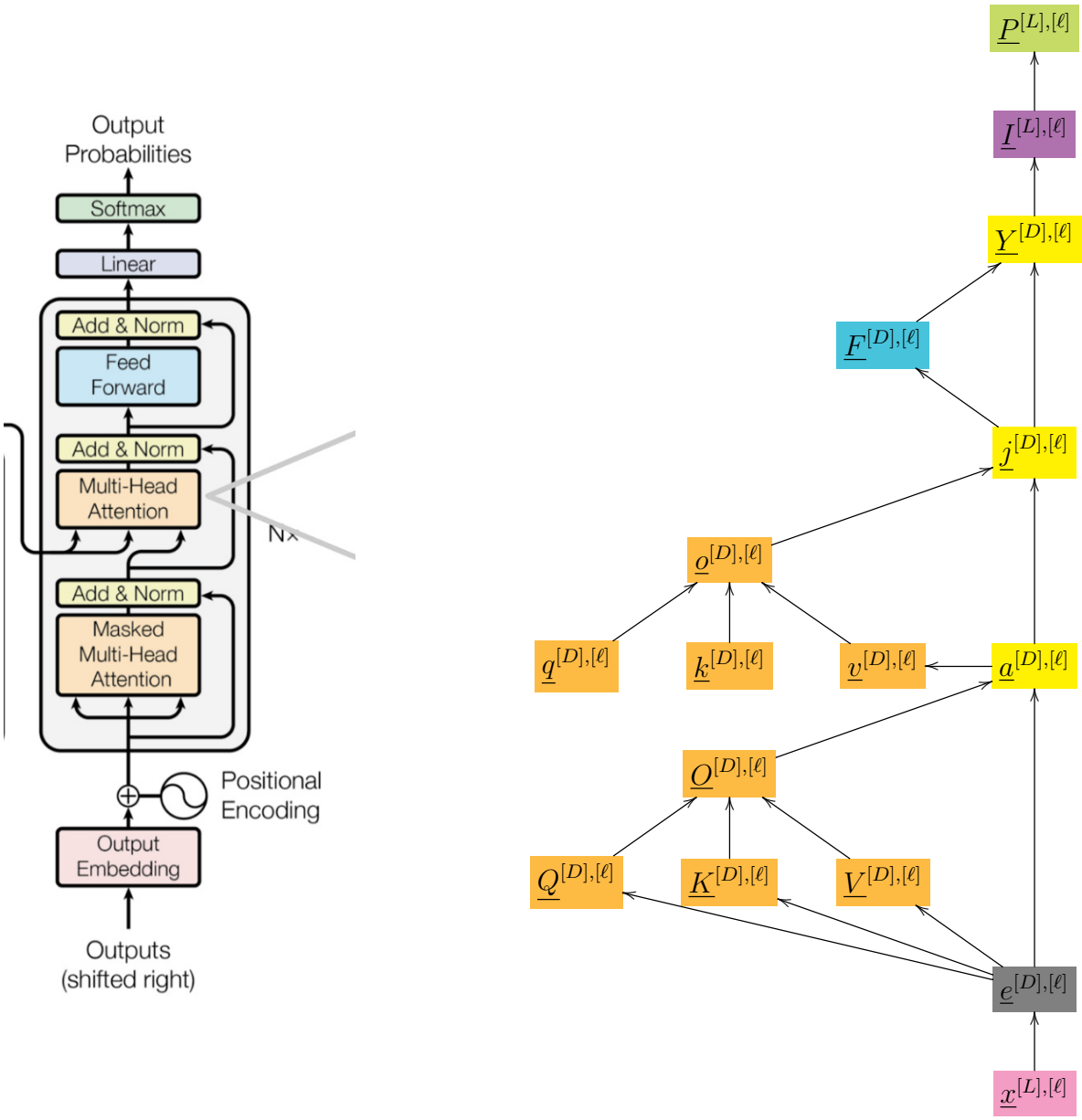


Figure 1: Decoder.

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

$$I^{[L],[\ell]} = W^{[L],[D]} Y^{[D],[\ell]} \tag{1b}$$

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

$$O^{[D],[\ell]} = \text{multi\_head\_attention}(Q^{[D],[\ell]}, K^{[D],[\ell]}, V^{[D],[\ell]}) \tag{1d}$$

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

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

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

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

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

$$e^{[D],[\ell]} = E^{[D],[L]} x^{[L],[\ell]} \tag{1j}$$

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

$$k^{[D],[\ell]} = \tag{1l}$$

$$o^{[D],[\ell]} = \text{multi\_head\_attention}(q^{[D],[\ell]}, k^{[D],[\ell]}, v^{[D],[\ell]}) \tag{1m}$$

$$q^{[D],[\ell]} = \tag{1n}$$

$$v^{[D],[\ell]} = a^{[D],[\ell]} \tag{1o}$$

$$x^{[L],[\ell]} = \tag{1p}$$