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

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

$$K^{[\Lambda],[D],[\ell]} = W_k^{[D],[d]} e^{[\Lambda],[d],[\ell]}$$
(1c)

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

$$(1d)$$

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

$$Q^{[\Lambda],[D],[\ell]} = W_q^{[D],[d]} e^{[\Lambda],[d],[\ell]}$$
(1f)

$$V^{[\Lambda],[D],[\ell]} = W_v^{[D],[d]} e^{[\Lambda],[d],[\ell]}$$
(1g)

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

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

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

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

$$\tag{1k}$$

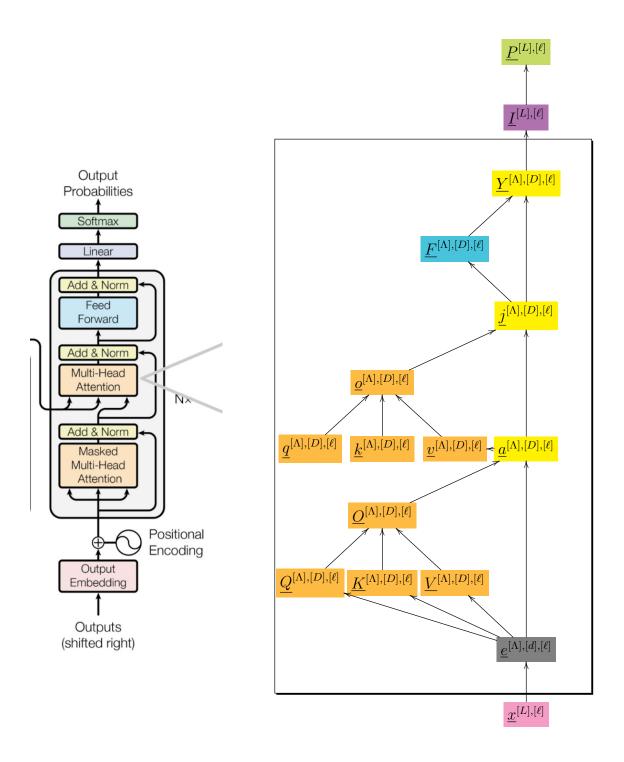
$$k^{[\Lambda],[D],[\ell]} = \text{prior, obtained from encoder.}$$
 (11)

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

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

$$v^{[\Lambda],[D],[\ell]} = a^{[\Lambda],[D],[\ell]}$$
 (10)

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



 \square Λ layers

Figure 1: Decoder.