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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

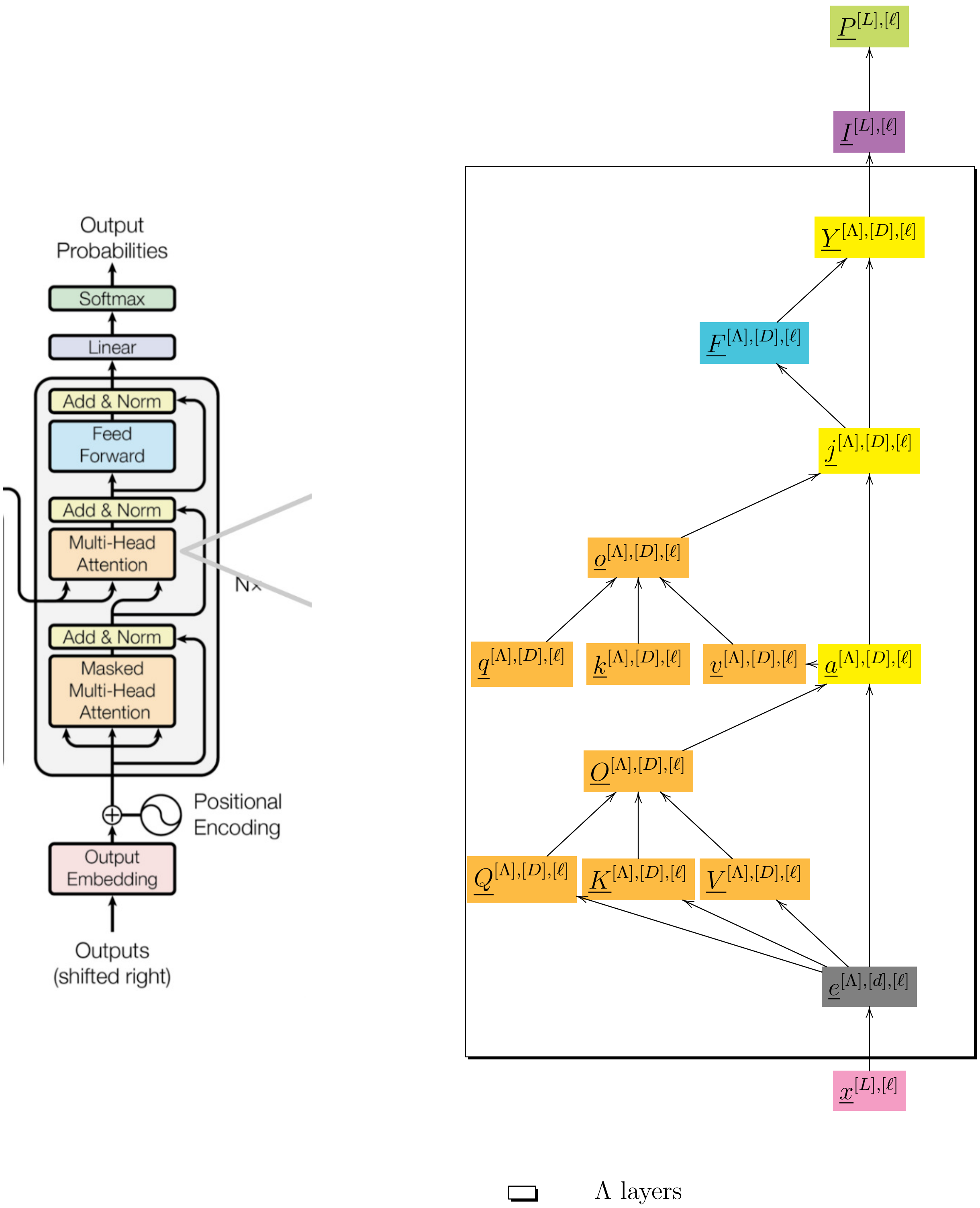
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
Add & Norm
Masked Multi-Head Attention
Positional Encoding
Output Embedding
Outputs (shifted right)
$\underline{P}^{[L],[\ell]}$
$\underline{I}^{[L],[\ell]}$
$\underline{Y}^{[\Lambda],[D],[\ell]}$
$\underline{F}^{[\Lambda],[D],[\ell]}$
$\underline{j}^{[\Lambda],[D],[\ell]}$
$\underline{o}^{[\Lambda],[D],[\ell]}$
$\underline{q}^{[\Lambda],[D],[\ell]}$
$\underline{k}^{[\Lambda],[D],[\ell]}$
$\underline{v}^{[\Lambda],[D],[\ell]}$
$\underline{a}^{[\Lambda],[D],[\ell]}$
$\underline{O}^{[\Lambda],[D],[\ell]}$
$\underline{Q}^{[\Lambda],[D],[\ell]}$
$\underline{K}^{[\Lambda],[D],[\ell]}$
$\underline{V}^{[\Lambda],[D],[\ell]}$
$\underline{e}^{[\Lambda],[d],[\ell]}$
$\underline{x}^{[L],[\ell]}$
▭ Λ layers

Figure 1: Decoder.