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

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

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

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

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

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

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

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

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

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

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

$$j^{[d],[\ell]} = \text{normalize}(U_{\underline{a}}^{[d],[D]} a^{[D],[\ell]} + J^{[d],[\ell]})$$
(11)

$$k^{[D],[\ell]} = U_k^{[D],[d]} n^{[d],[\ell]}$$
 (1m)

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

$$q^{[D],[\ell]} = U_{\underline{q}}^{[D],[d]} J^{[d],[\ell]}$$
 (10)

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

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

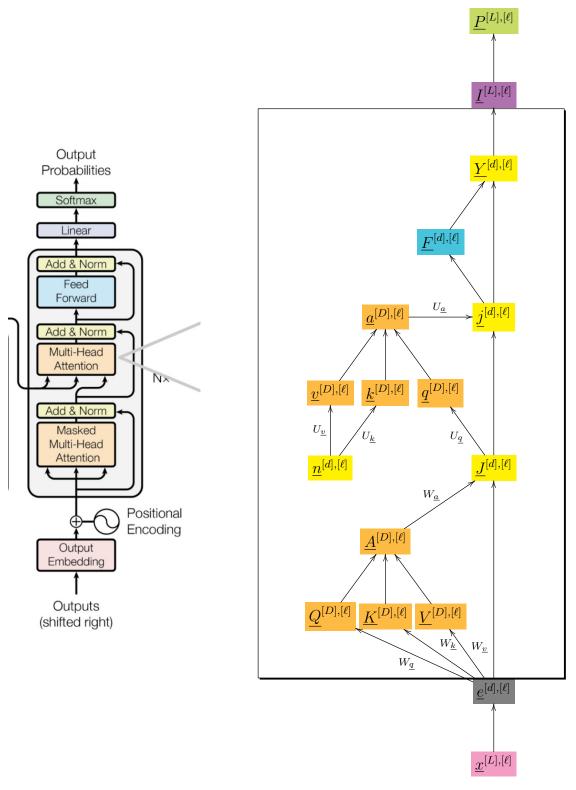


Figure 1: Decoder of Vanilla Transformer Net. Λ copies of the boxed part are connected in series.