



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

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

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

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

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

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

$$j^{[d],[\ell]} = \text{normalize}(U_{\underline{a}}^{[d],[D]} a^{[D],[\ell]} + J^{[d],[\ell]}) \quad (1f)$$

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

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

$$k^{[D],[\ell]} = U_{\underline{k}}^{[D],[d]} n^{[d],[\ell]} \quad (1i)$$

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

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

$$q^{[D],[\ell]} = U_{\underline{q}}^{[D],[d]} J^{[d],[\ell]} \quad (1l)$$

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

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

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

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

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