



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

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

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

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

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

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

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

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

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

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

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

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

$$v^{[D],[\ell]} = a^{[D],[\ell]} \quad (1o)$$

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