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)

$\underline{G}^{[D],[L]}$

$\underline{I}^{[D],[L]}$

$\underline{Y}^{[D],[L]}$

$\underline{F}^{[D],[L]}$

$\underline{j}^{[D],[L]}$

$\underline{o}^{[L]}$

$\underline{q}^{[D],[L]}$

$\underline{k}^{[D],[L]}$

$\underline{v}^{[D],[L]}$

$\underline{a}^{[D],[L]}$

$\underline{Q}^{[L]}$

$\underline{Q}^{[D],[L]}$

$\underline{K}^{[D],[L]}$

$\underline{V}^{[D],[L]}$

$\underline{p}^{[D],[L]}$

$\underline{R}^{[D],[L]}$

Figure 1: Decoder.

$$G^{[D],[L]} = I^{[D],[L]} \tag{1a}$$

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

1

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

$$F^{[D],[L]} = j^{[D],[L]} \tag{1d}$$

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

$$o^{[L]} = q^{[D],[L]}, k^{[D],[L]}, v^{[D],[L]} \tag{1f}$$

$$a^{[D],[L]} = \text{normalize}(O^{[L]} + p^{[D],[L]}) \tag{1g}$$

$$v^{[D],[L]} = a^{[D],[L]} \tag{1h}$$

$$O^{[L]} = Q^{[D],[L]}, K^{[D],[L]}, V^{[D],[L]} \tag{1i}$$

$$Q^{[D],[L]} = p^{[D],[L]} \tag{1j}$$

$$K^{[D],[L]} = p^{[D],[L]} \tag{1k}$$

$$V^{[D],[L]} = p^{[D],[L]} \tag{1l}$$

$$p^{[D],[L]} = R^{[D],[L]} \tag{1m}$$

$$R^{[D],[L]} = \tag{1n}$$

$$q^{[D],[L]} = \tag{1o}$$

2

$$k^{[D],[L]} = \tag{1p}$$

3