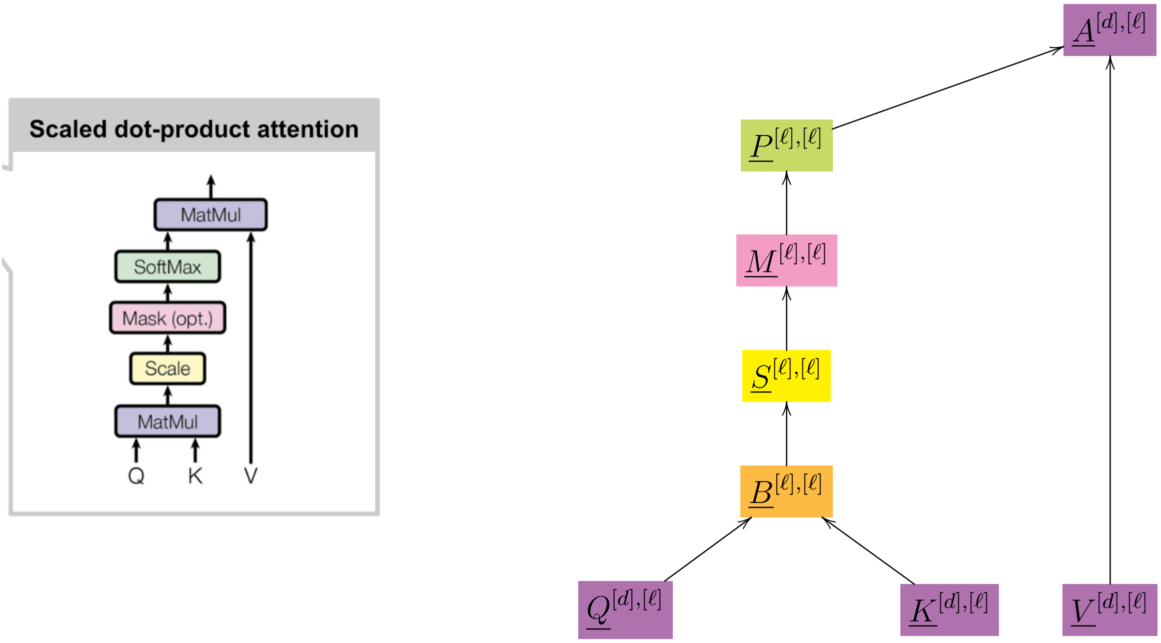
Scaled dot-product attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
$\underline{A}^{[d],[\ell]}$
$\underline{P}^{[\ell],[\ell]}$
$\underline{M}^{[\ell],[\ell]}$
$\underline{S}^{[\ell],[\ell]}$
$\underline{B}^{[\ell],[\ell]}$
$\underline{Q}^{[d],[\ell]}$
$\underline{K}^{[d],[\ell]}$
$\underline{V}^{[d],[\ell]}$

Figure 1: Scaled Dot Product Attention.

$$A^{[d],[\ell]} = V^{[d],[\ell]} P^{[\ell],[\ell]} \tag{1a}$$

$$B^{[\ell],[\ell]} = (Q^{[d],[\ell]})^T K^{[d],[\ell]} \tag{1b}$$

$$K^{[d],[\ell]} = prior \tag{1c}$$

$$M^{[\ell],[\ell]} = \text{mask}(S^{[\ell],[\ell]}) \tag{1d}$$

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

$$Q^{[d],[\ell]} = prior \tag{1f}$$

$$S^{[\ell],[\ell]} = \frac{B^{[\ell],[\ell]}}{\sqrt{d}} \tag{1g}$$

$$V^{[d],[\ell]} = prior \tag{1h}$$