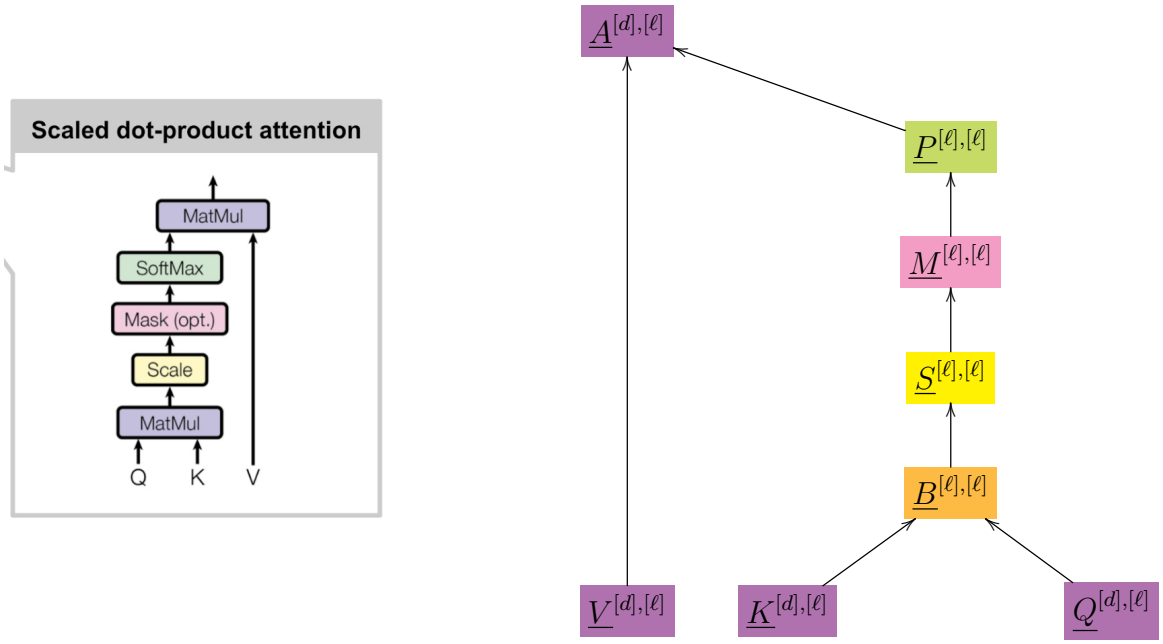

Figure 1: Scaled Dot Product Attention.

$$A^{[d],[\ell]} = V^{[d],[\ell]} P^{[\ell],[\ell]} \quad \left(\text{Note that } \sum_{\alpha \in [\ell]} P^{\alpha,[\ell]} = 1 \right) \tag{1a}$$

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

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

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

$$P^{[\ell],[\ell]} = \text{softmax}(M^{[\ell],[\ell]}) \quad \left(\text{Note that } \sum_{\alpha \in [\ell]} P^{\alpha,[\ell]} = 1 \right) \tag{1e}$$

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

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

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