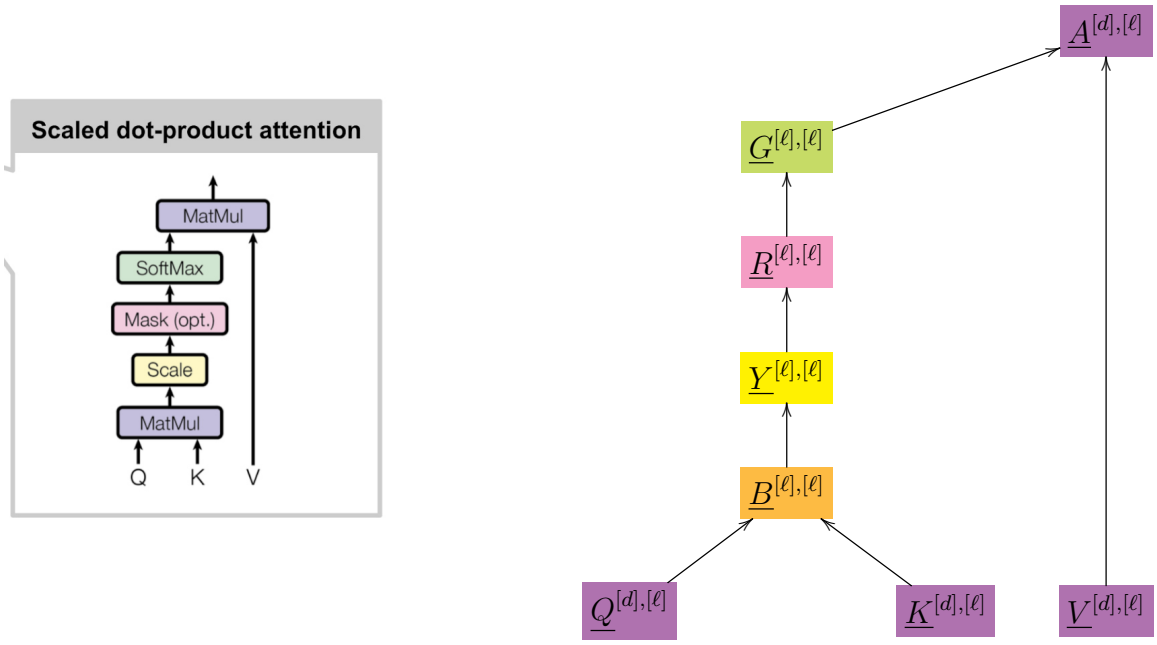

Figure 1: Scaled Dot Product Attention.

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

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

$$G^{[\ell],[\ell]} = \text{softmax}(R^{[\ell],[\ell]}) \tag{1c}$$

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

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

$$R^{[\ell],[\ell]} = \text{mask}(Y^{[\ell],[\ell]}) \tag{1f}$$

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

$$Y^{[\ell],[\ell]} = \frac{B^{[\ell],[\ell]}}{\sqrt{d_{\underline{k}}}} \tag{1h}$$