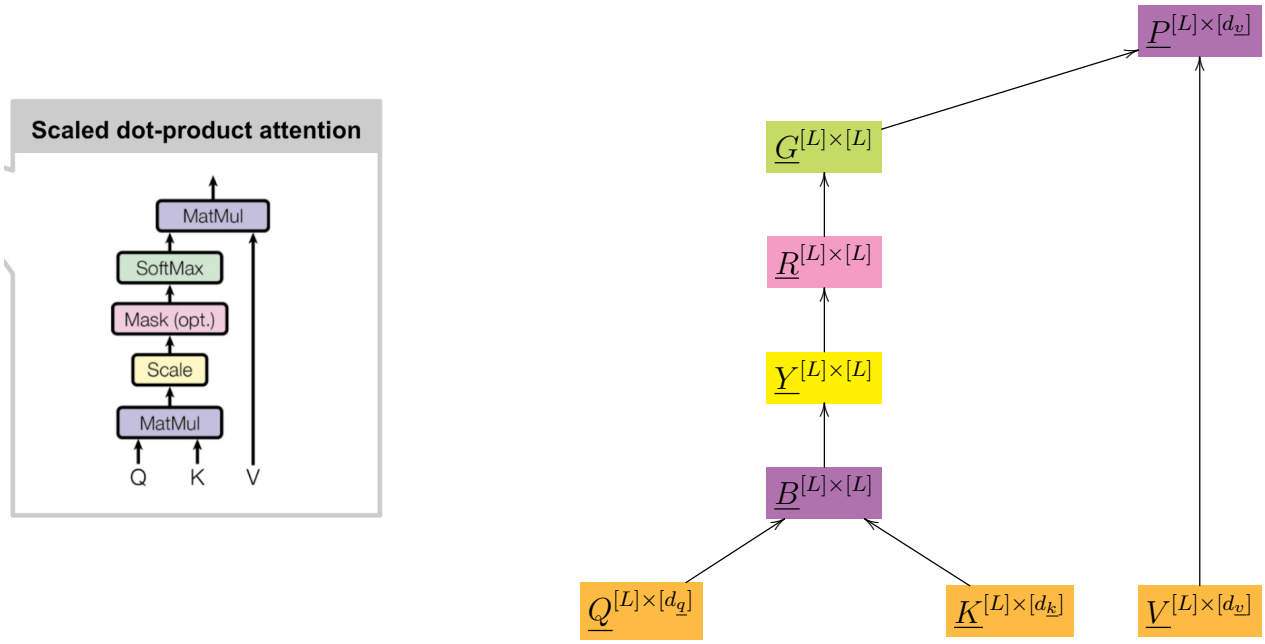

Figure 1: Scaled Dot Product Attention.

$$Q^{[L]\times[d_q]} = prior \tag{1a}$$

$$K^{[L]\times[d_k]} = prior \tag{1b}$$

$$V^{[L]\times[d_v]} = prior \tag{1c}$$

$$B^{[L]\times[L]} = Q^{[L]\times[d_q]}(K^{[L]\times[d_k]})^{\dagger} \tag{1d}$$

$$Y^{[L]\times[L]} = \frac{B^{[L]\times[L]}}{\sqrt{d_k}} \tag{1e}$$

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

$$G^{[L]\times[L]} = \text{softmax}(R^{[L]\times[L]}) \tag{1g}$$

$$P^{[L]\times[d_v]} = G^{[L]\times[L]}V^{[L]\times[d_v]} \tag{1h}$$