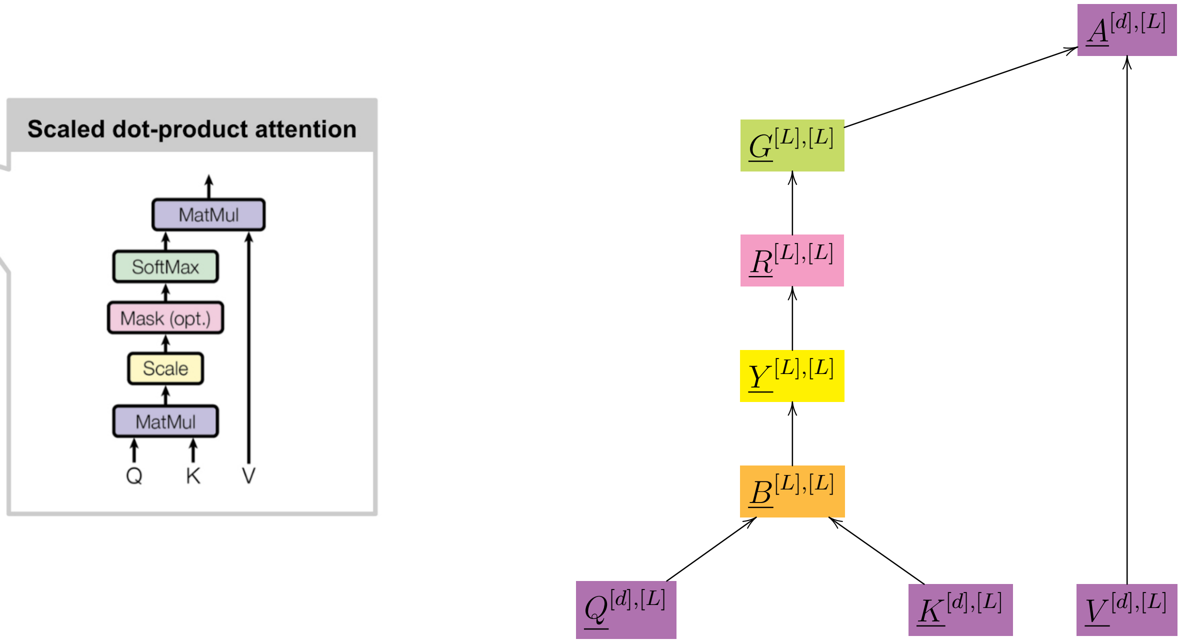
Scaled dot-product attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
$\underline{A}^{[d],[L]}$
$\underline{G}^{[L],[L]}$
$\underline{R}^{[L],[L]}$
$\underline{Y}^{[L],[L]}$
$\underline{B}^{[L],[L]}$
$\underline{Q}^{[d],[L]}$
$\underline{K}^{[d],[L]}$
$\underline{V}^{[d],[L]}$

Figure 1: Scaled Dot Product Attention.

$$Q^{[d],[L]} =; prior \tag{1a}$$

$$K^{[d],[L]} =; prior \tag{1b}$$

$$V^{[d],[L]} =; prior \tag{1c}$$

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

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

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

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

$$A^{[d],[L]} = V^{[d],[L]} G^{[L],[L]} \tag{1h}$$