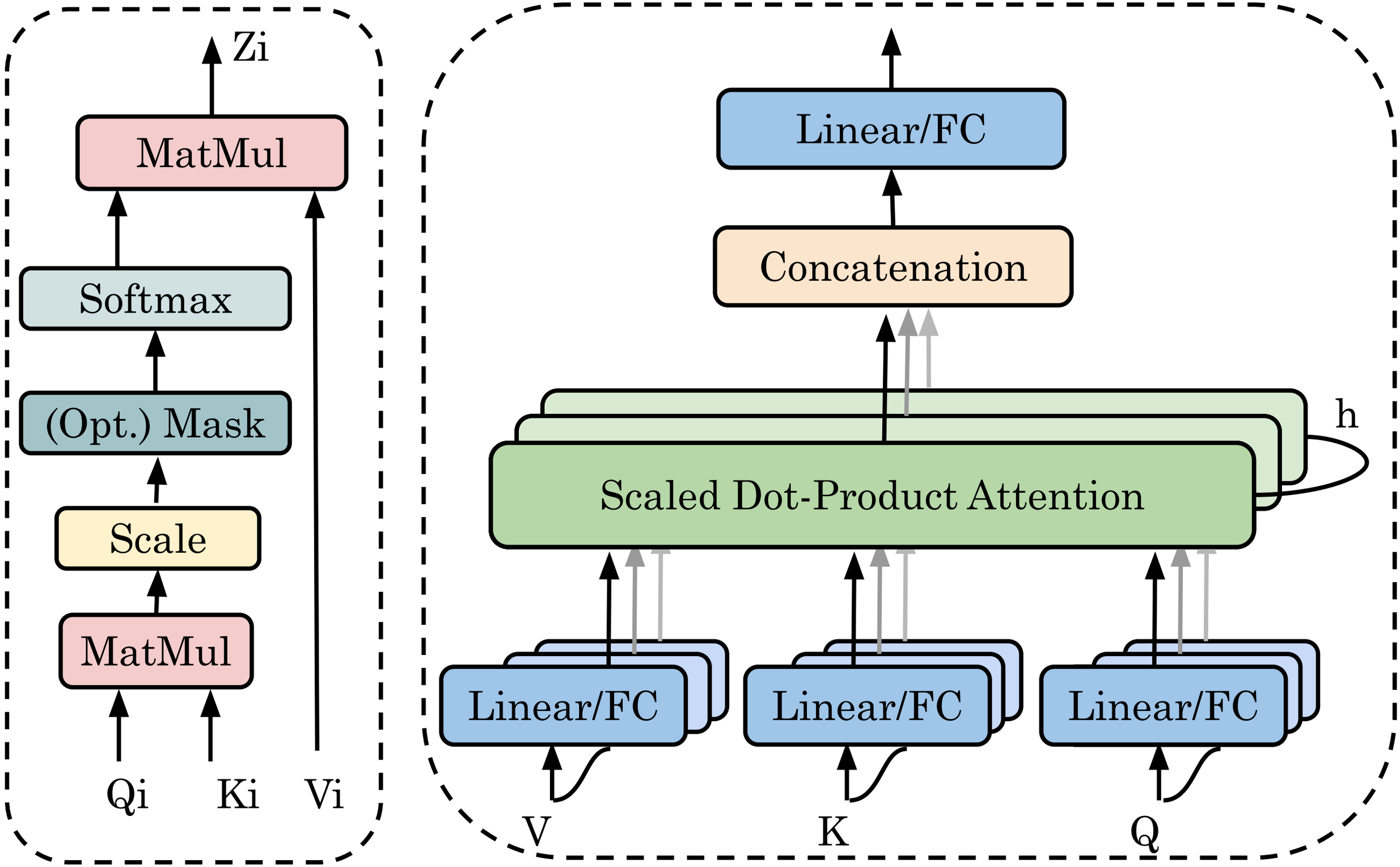

(a) Scaled Dot-product Attention

(b) Multi-Head Self-Attention