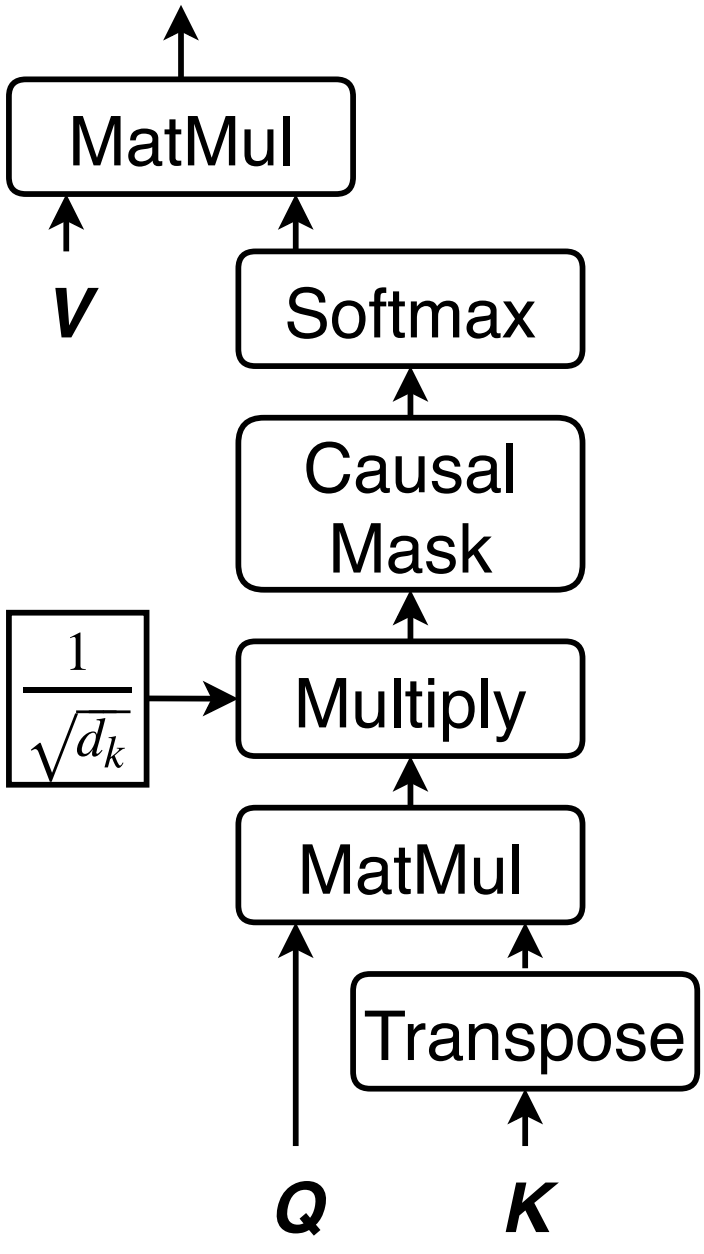
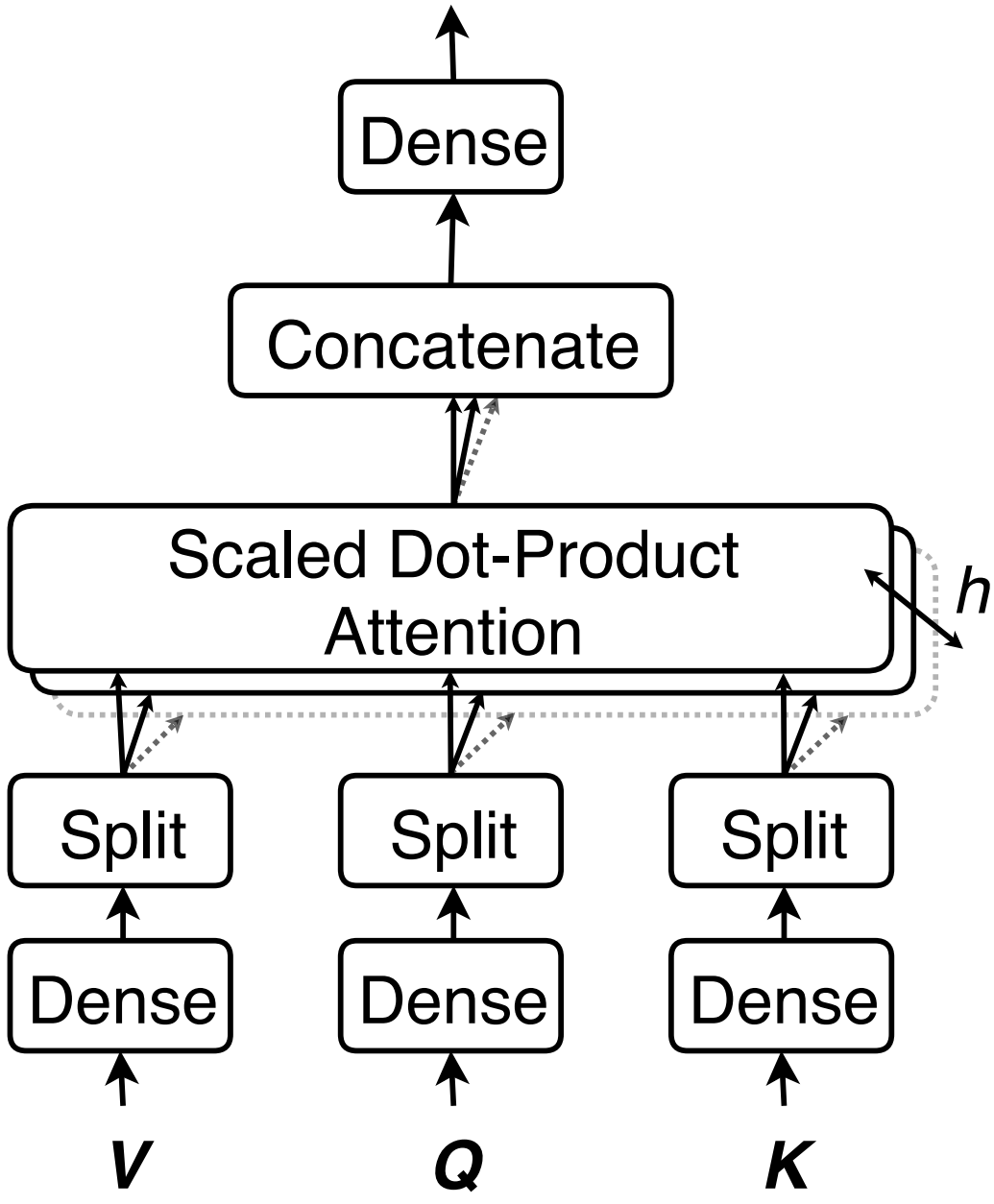


(a) Scaled Dot-Product Attention



(b) Multi-Head Attention