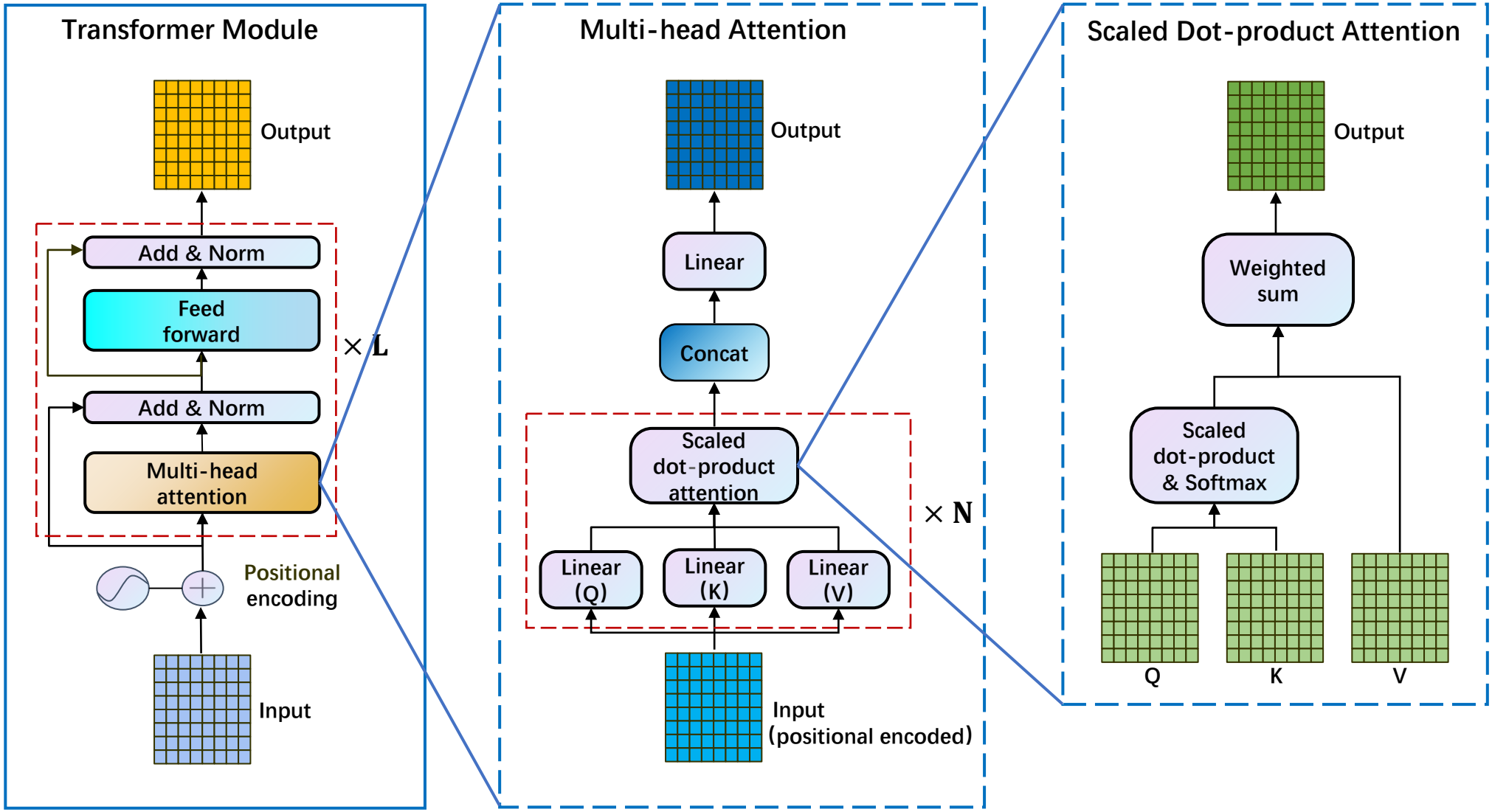
Transformer Module
Output
Add & Norm
Feed forward
× L
Add & Norm
Multi-head attention
Positional encoding
Input
Multi-head Attention
Output
Linear
Concat
Scaled dot-product attention
× N
Linear (Q)
Linear (K)
Linear (V)
Input (positional encoded)
Scaled Dot-product Attention
Output
Weighted sum
Scaled dot-product & Softmax
Q
K
V