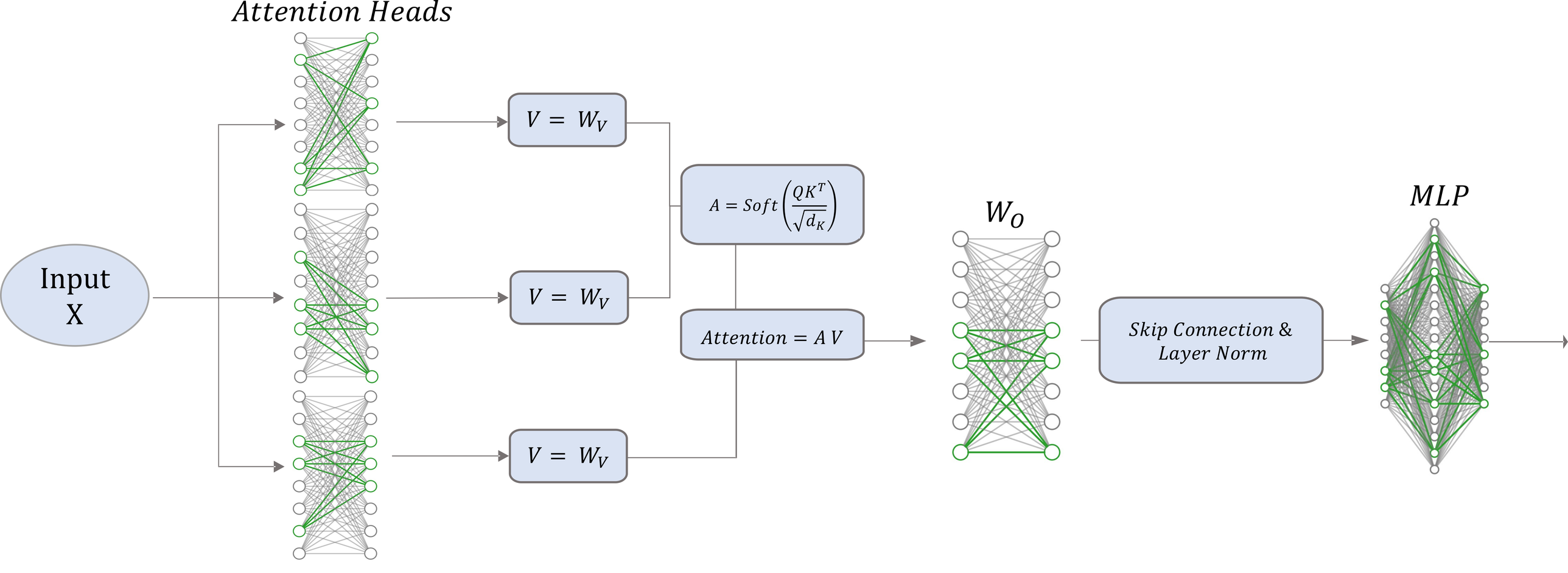
Attention Heads
Input
X
$V = W_V$
$V = W_V$
$V = W_V$
$A = Soft\left(\frac{QK^T}{\sqrt{d_K}}\right)$
$Attention = A\,V$
W_O
Skip Connection &
Layer Norm
MLP