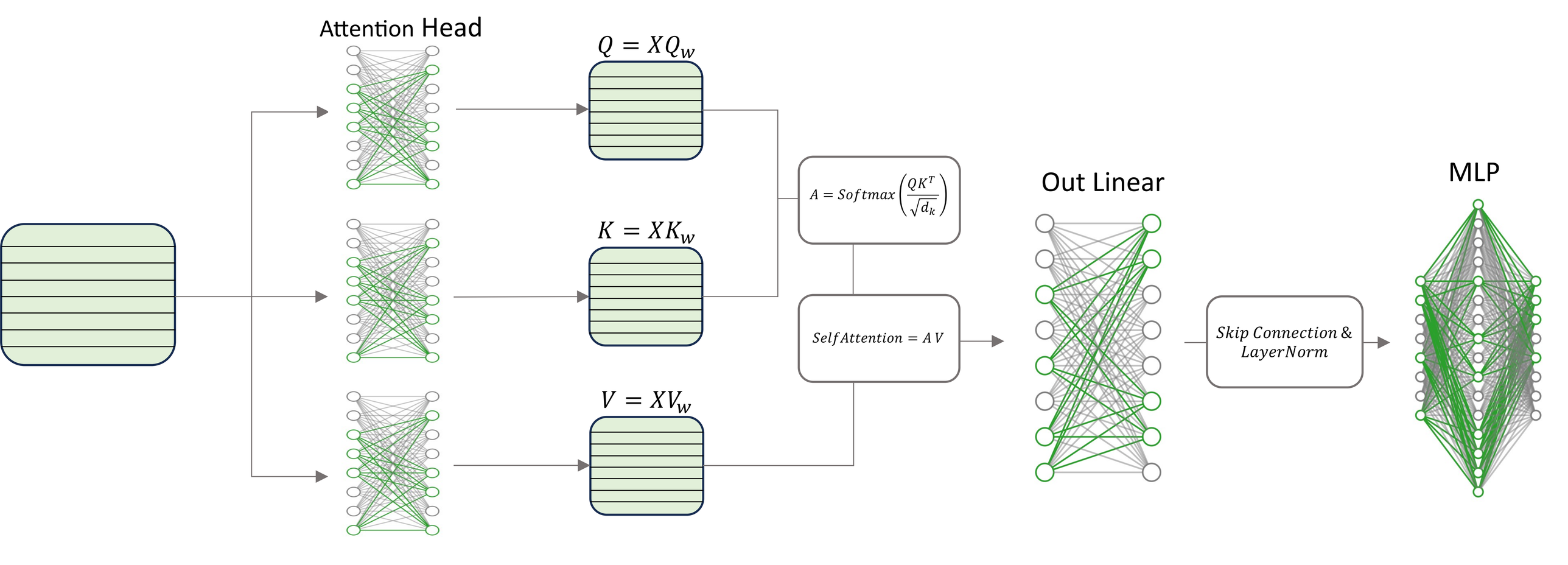
Attention Head
$Q = XQ_w$
$K = XK_w$
$V = XV_w$
$A = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$
$SelfAttention = A\,V$
Out Linear
$Skip\ Connection\ \&\ LayerNorm$
MLP