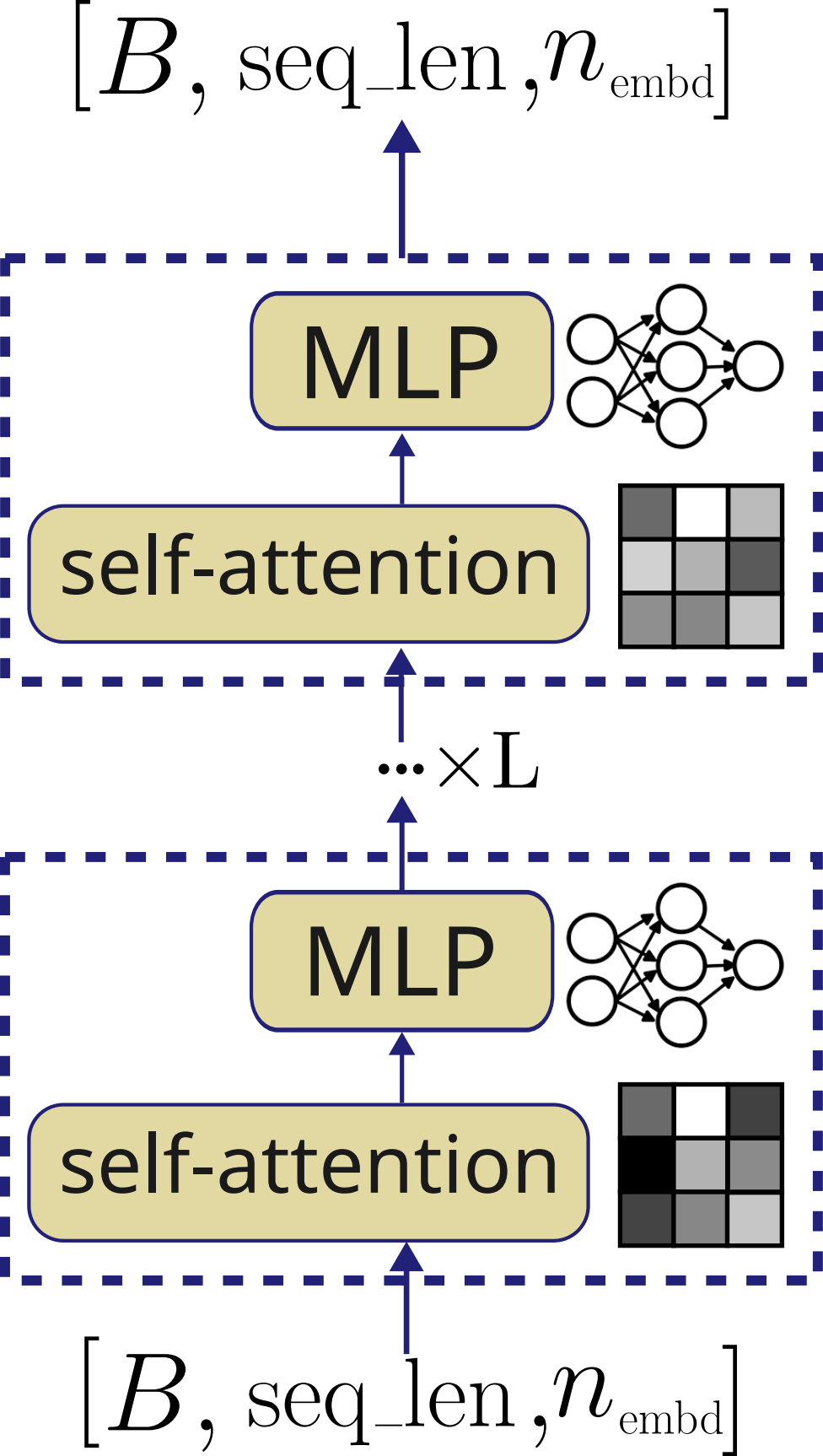
$[B, \text{seq\_len}, n_{\text{embd}}]$
MLP
self-attention
$\cdots \times \text{L}$
MLP
self-attention
$[B, \text{seq\_len}, n_{\text{embd}}]$