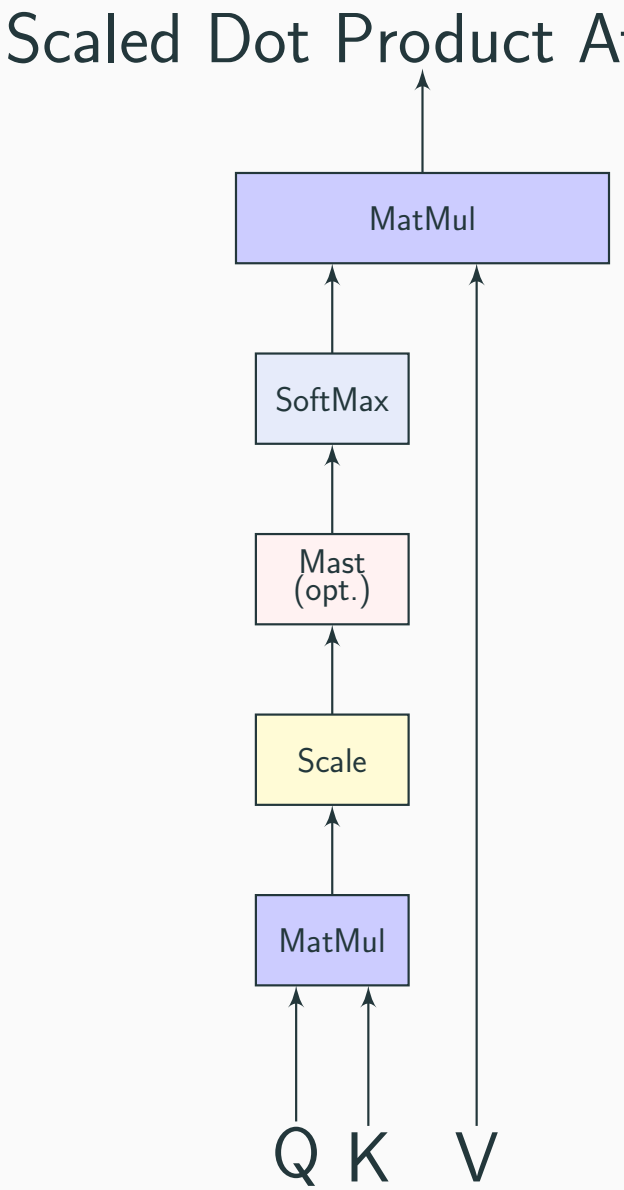
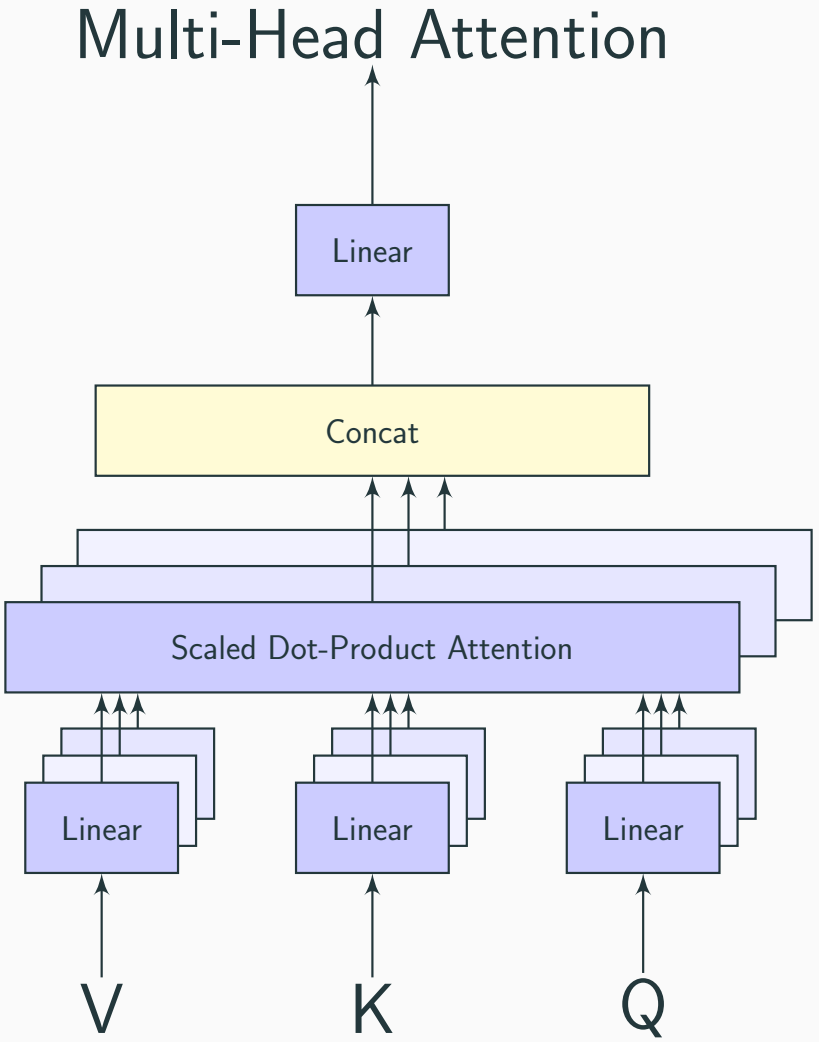
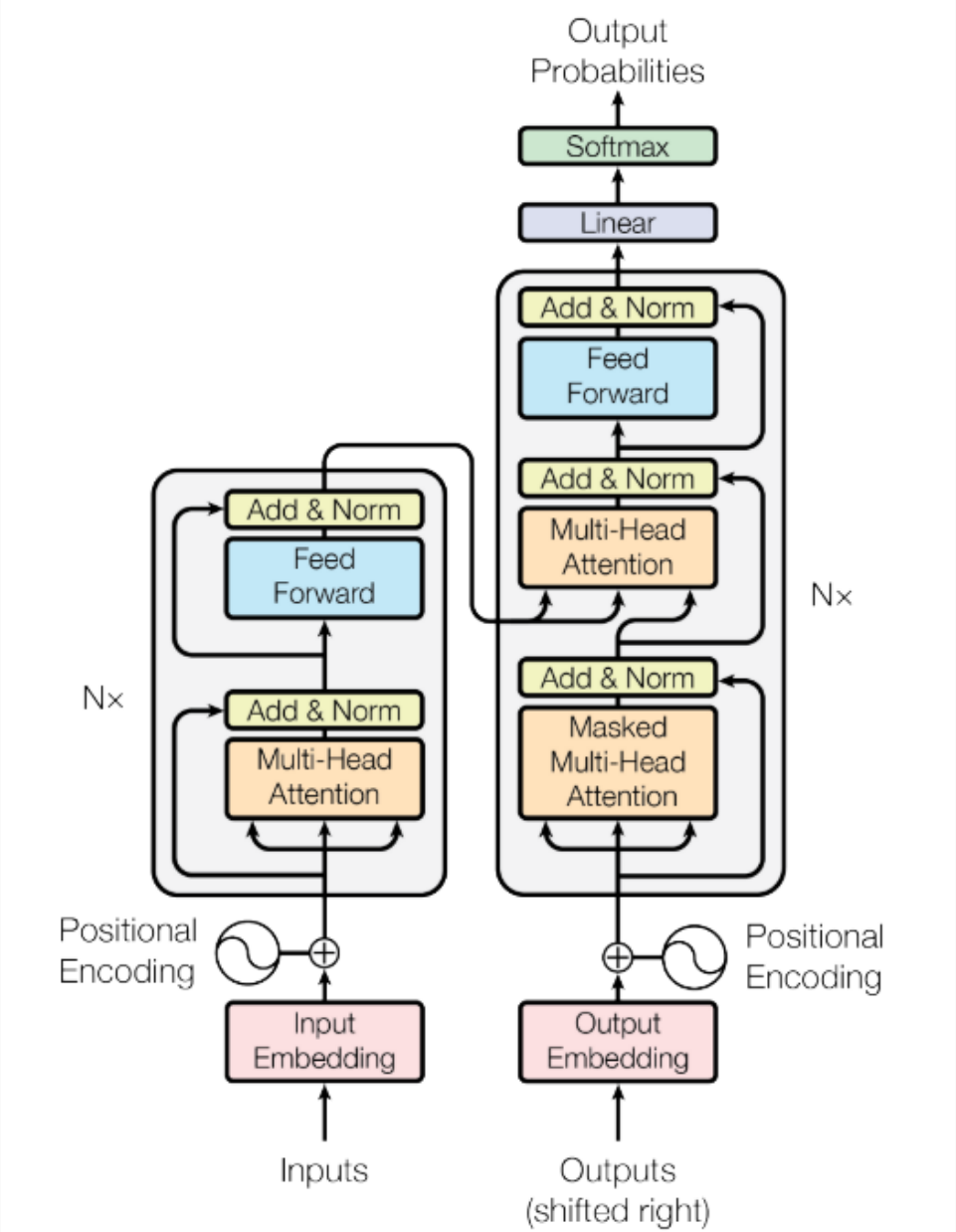
Scaled Dot product



Multi-Head Attention



The Transformer - model architecture



Positional Encoding - Example

$$PE_{(pos,2_i)} = sin(pos/1000^{2i/d_{model}})$$

$$PE_{(pos,2_{i+1})} = cos(pos/1000^{2i/d_{model}})$$