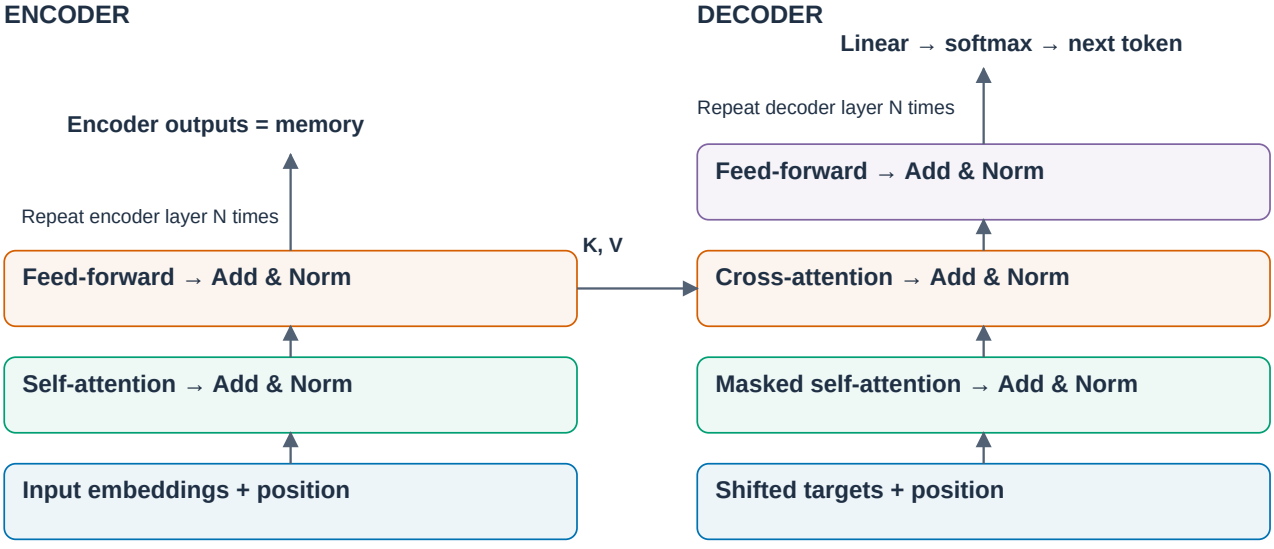
Transformer: encoder memory meets causal decoding

Classic post-norm architecture (Vaswani et al., 2017); arrows run bottom → top.



Decoder Q attends to encoder K,V; causal masking prevents access to future target tokens.

Original educational redraw · schematic values only · SVG text and objects remain editable