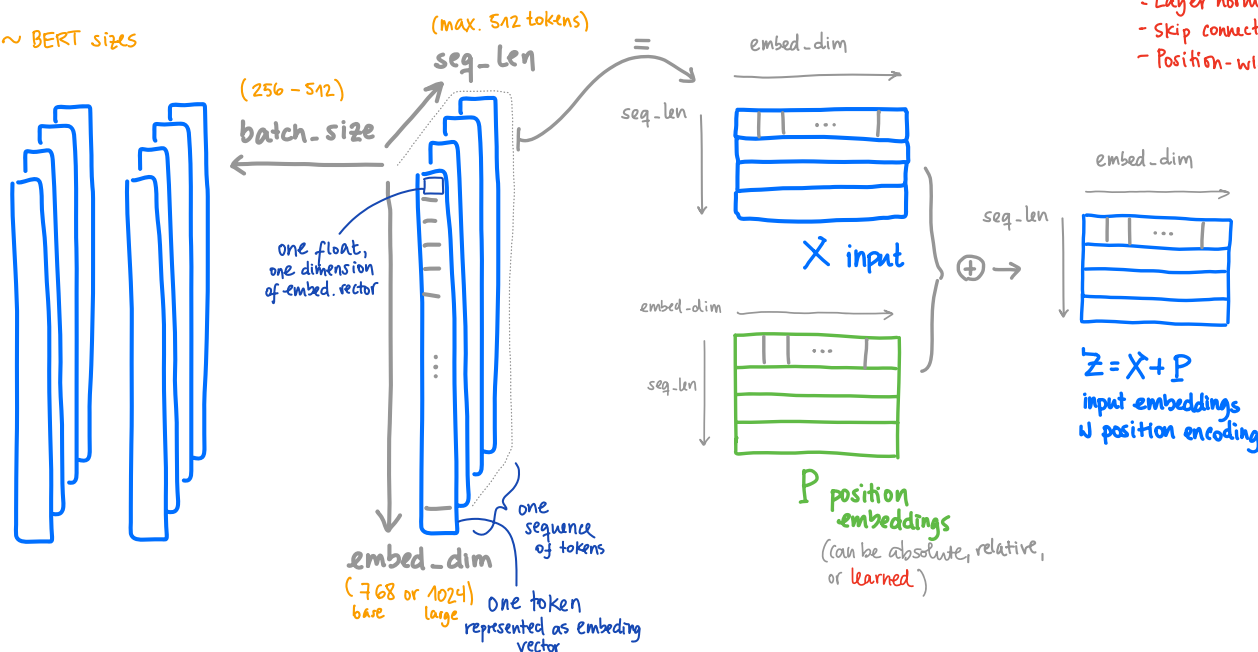


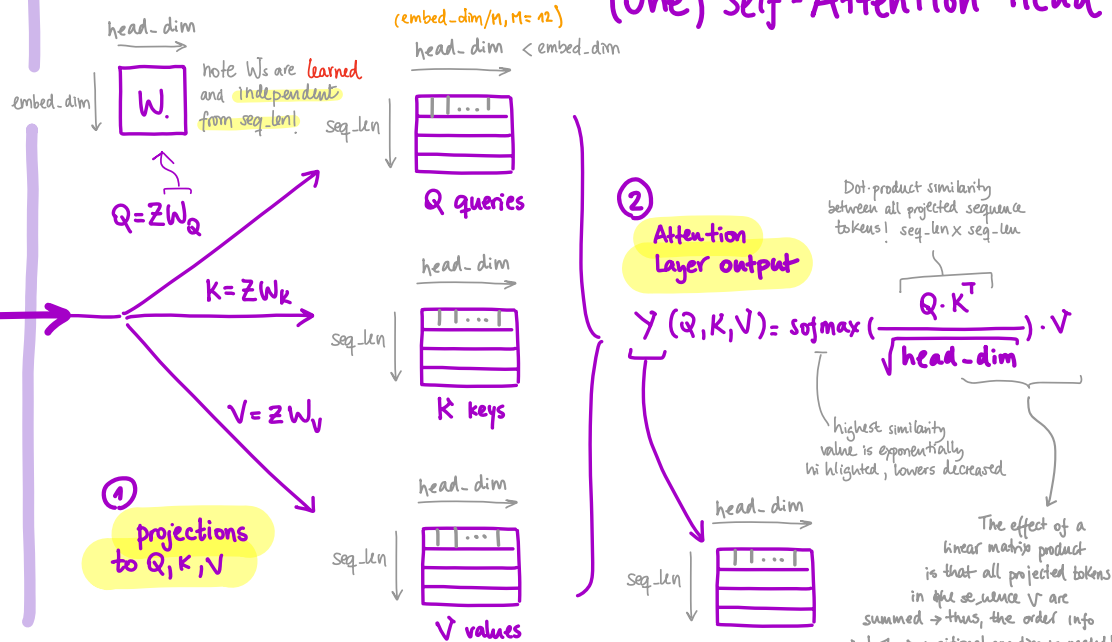
SIMPLIFIED TRANSFORMER-ENCODER ARCHITECTURE

- ⚠ Missing components:
- Layer normalization
 - Skip connections
 - Position-wise feed-forward layer

~ BERT sizes

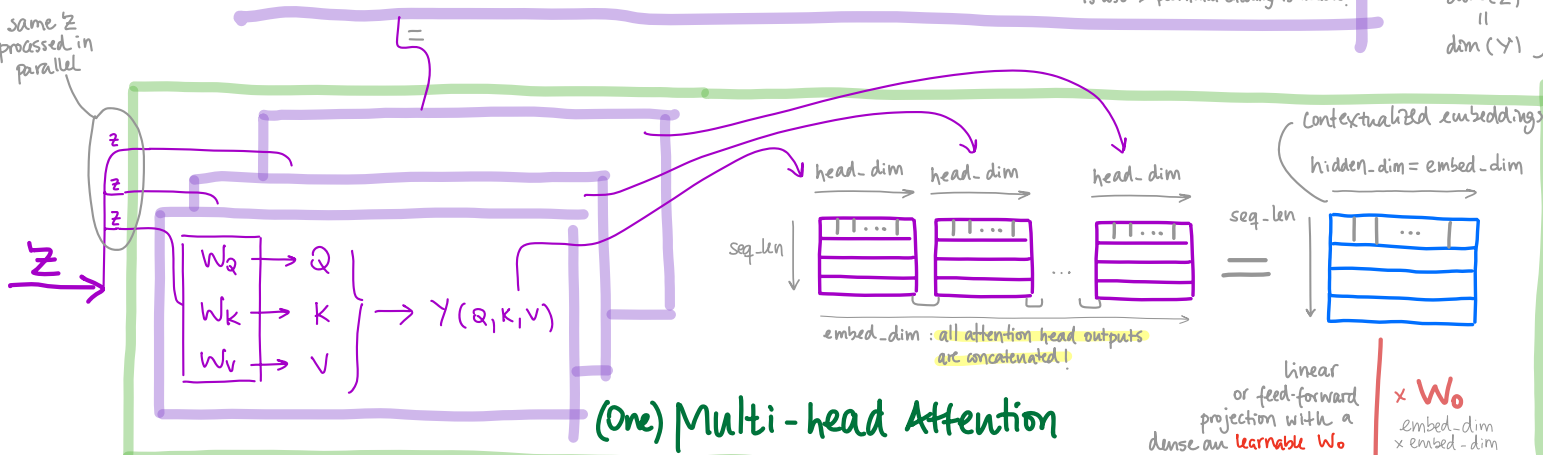


(One) Self-Attention Head



$\dim(X)$
||
 $\dim(Z)$
||
 $\dim(Y)$

(One) Multi-head Attention



→ We have N multi-head attention blocks stacked serially ($N = 12$ (base) or 16 (large) in BERT)

(M self-attention layers)
($M = 12$ in BERT)

Linear or feed-forward projection with a dense and learnable W_O
 $\times W_O$
 $\text{embed-dim} \times \text{embed-dim}$