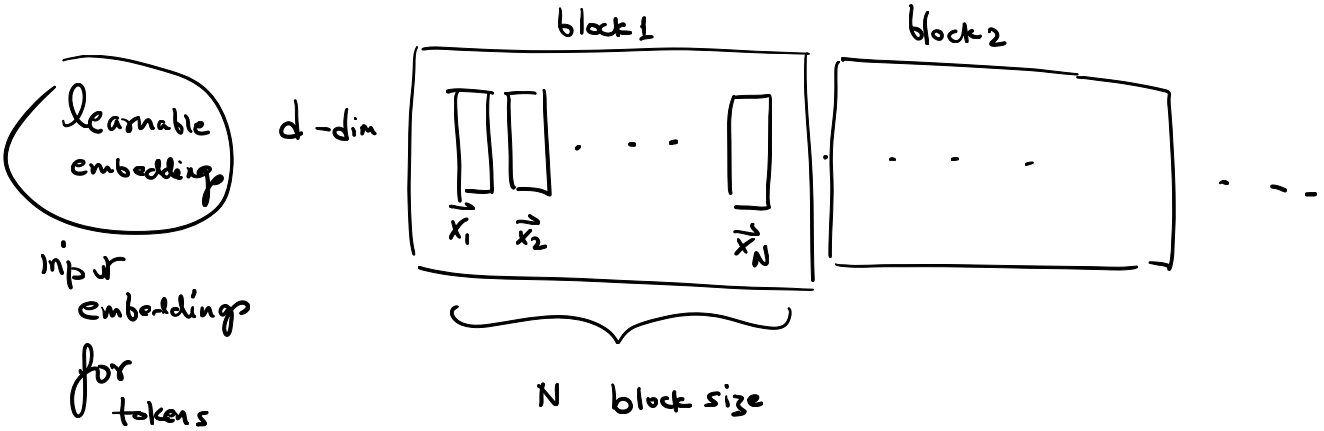
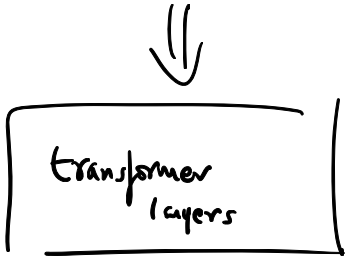
Transformers

dot-product attention

Input: block/"sequence"

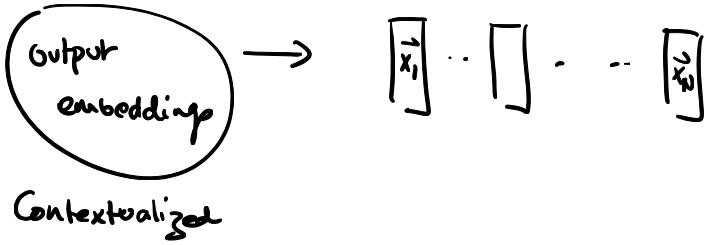


Cost per block $O(N^2)$



Same token

different embeddings in diff blocks



Self-attention