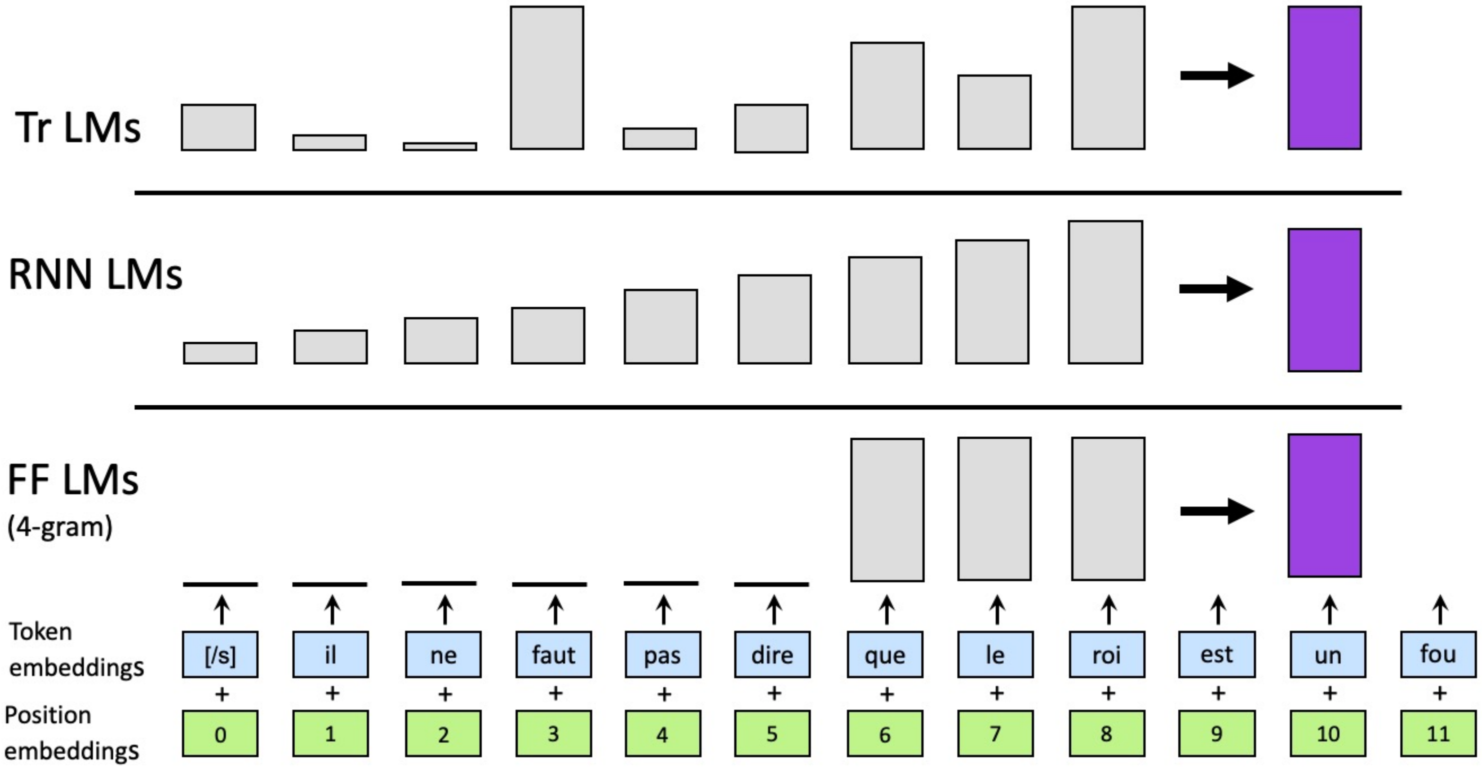
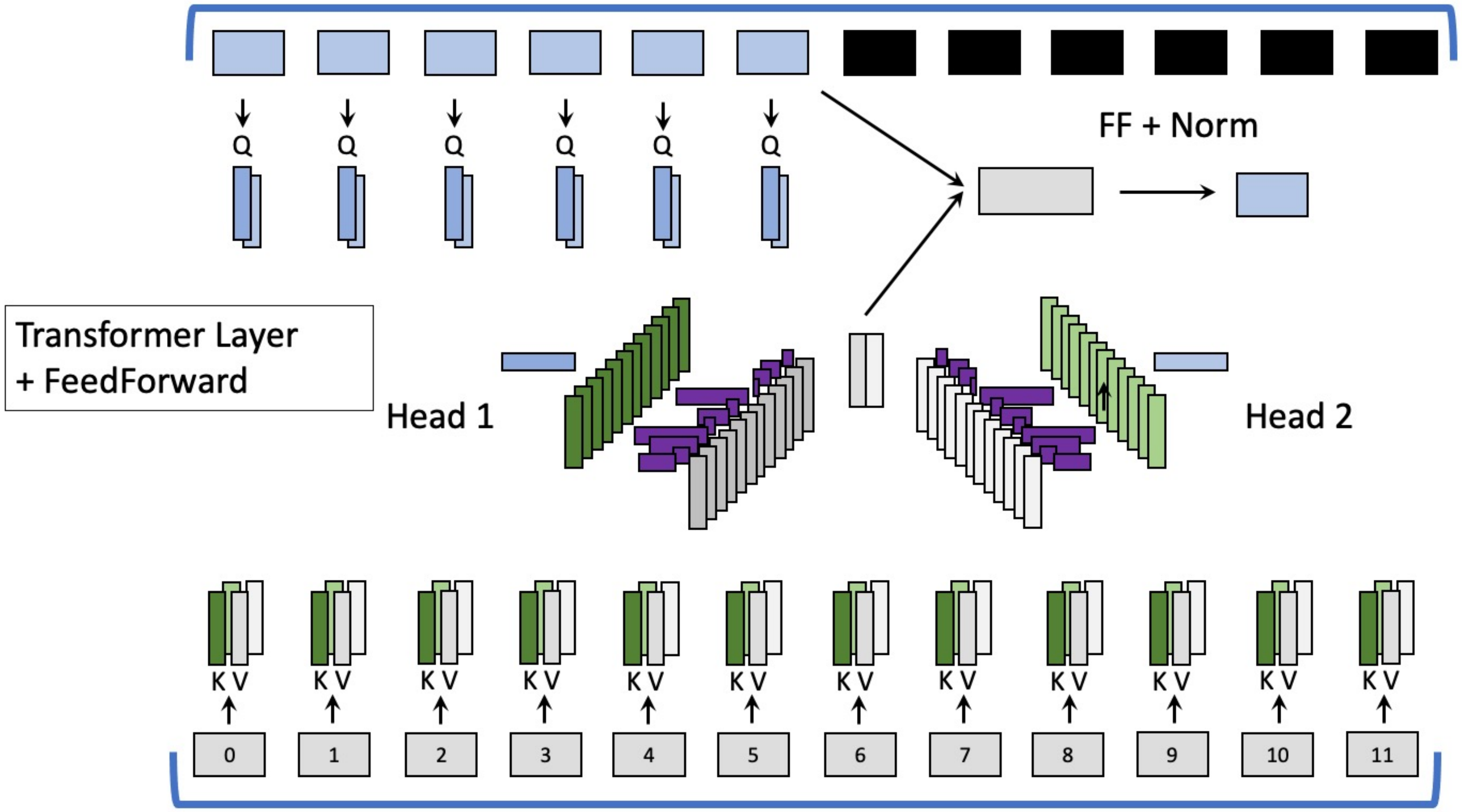
Transformers as pure LMs: improving the past context



Also: relax causality, recompute past representations after each new word

Faster, Better Encoder-Decoder + Attention : Transformer



In the decoder: cross-attention with respect to the last encoder layer