

The Transformer architecture

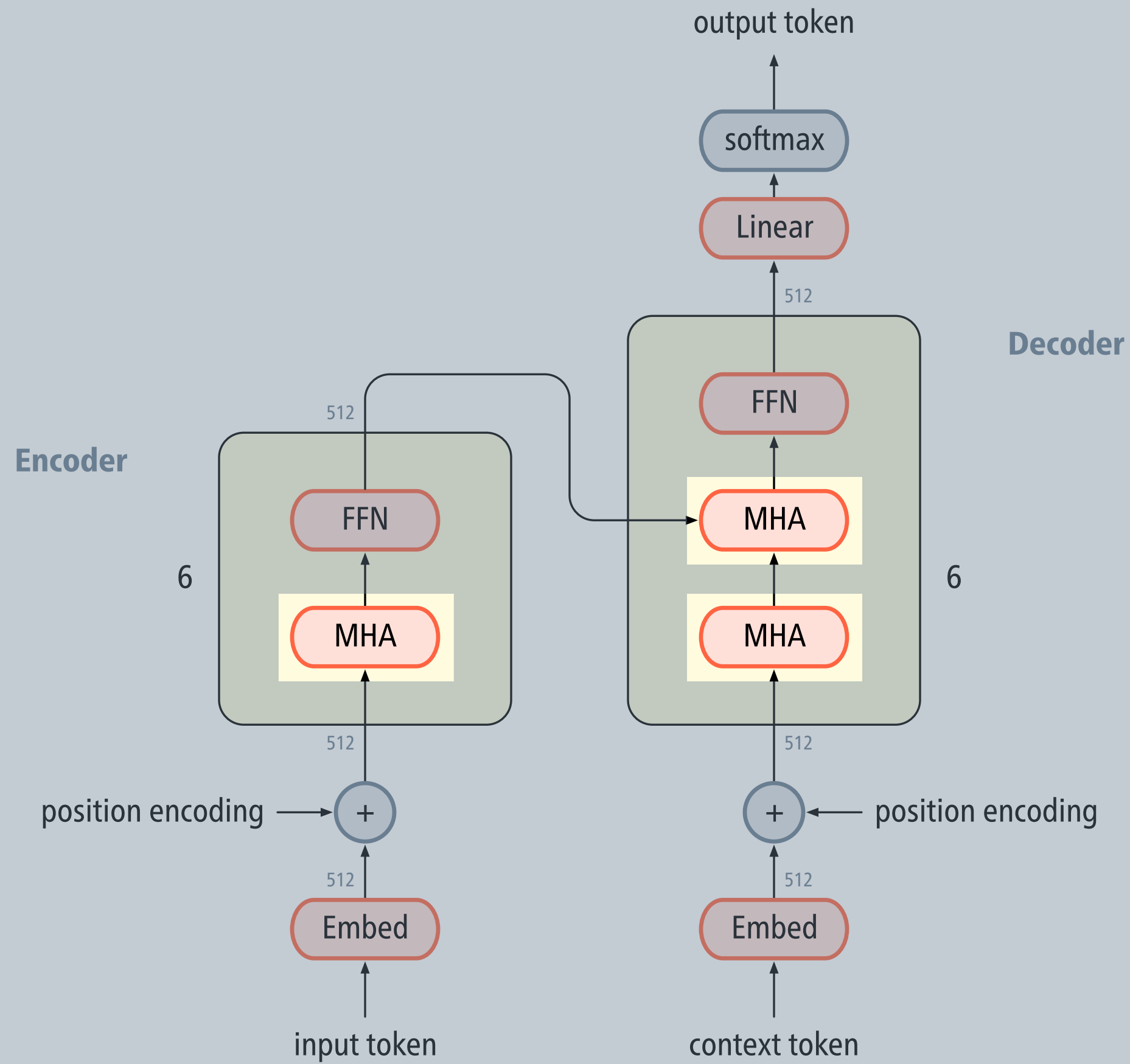
Marco Kuhlmann

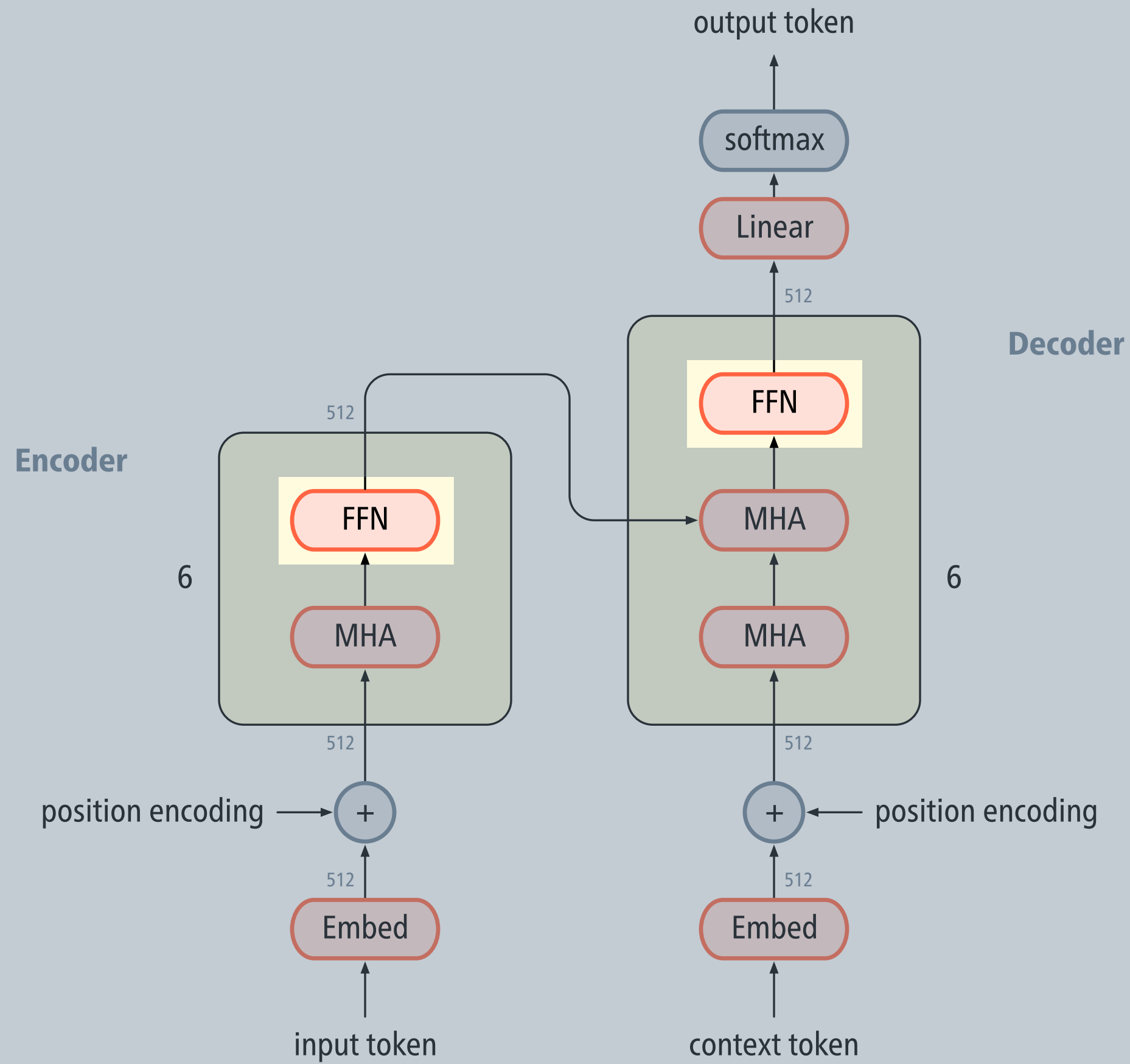
Department of Computer and Information Science

Attention is all you need

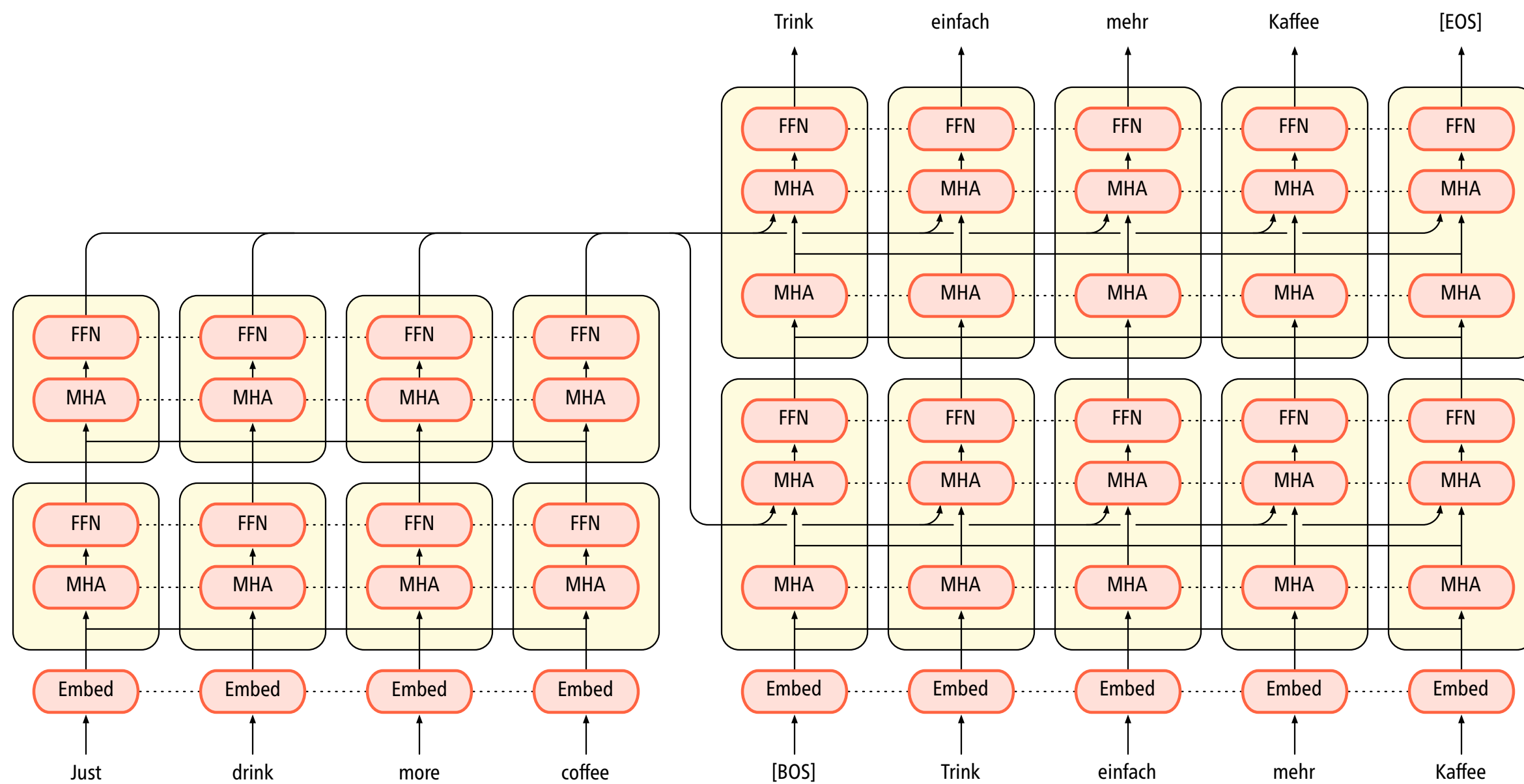
- Recurrent neural networks implement a sequential model of computation that processes sequences element-by-element.
- In contrast, attention facilitates direct access to all elements, independently of sequence length.
- The **Transformer** is an encoder–decoder architecture that drops recurrent neural networks and exclusively uses attention.

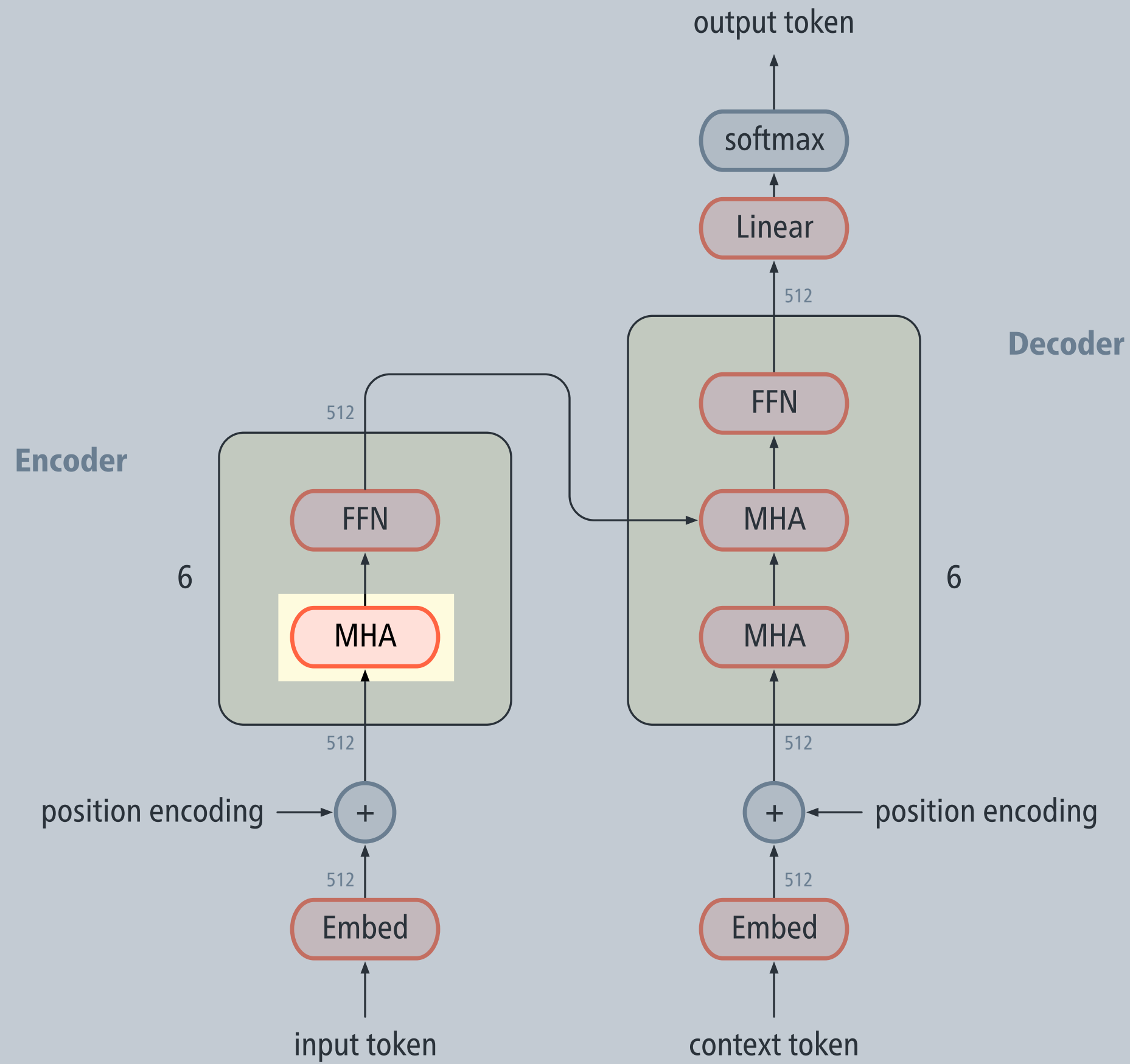
fixes information bottleneck problem; can be parallelised



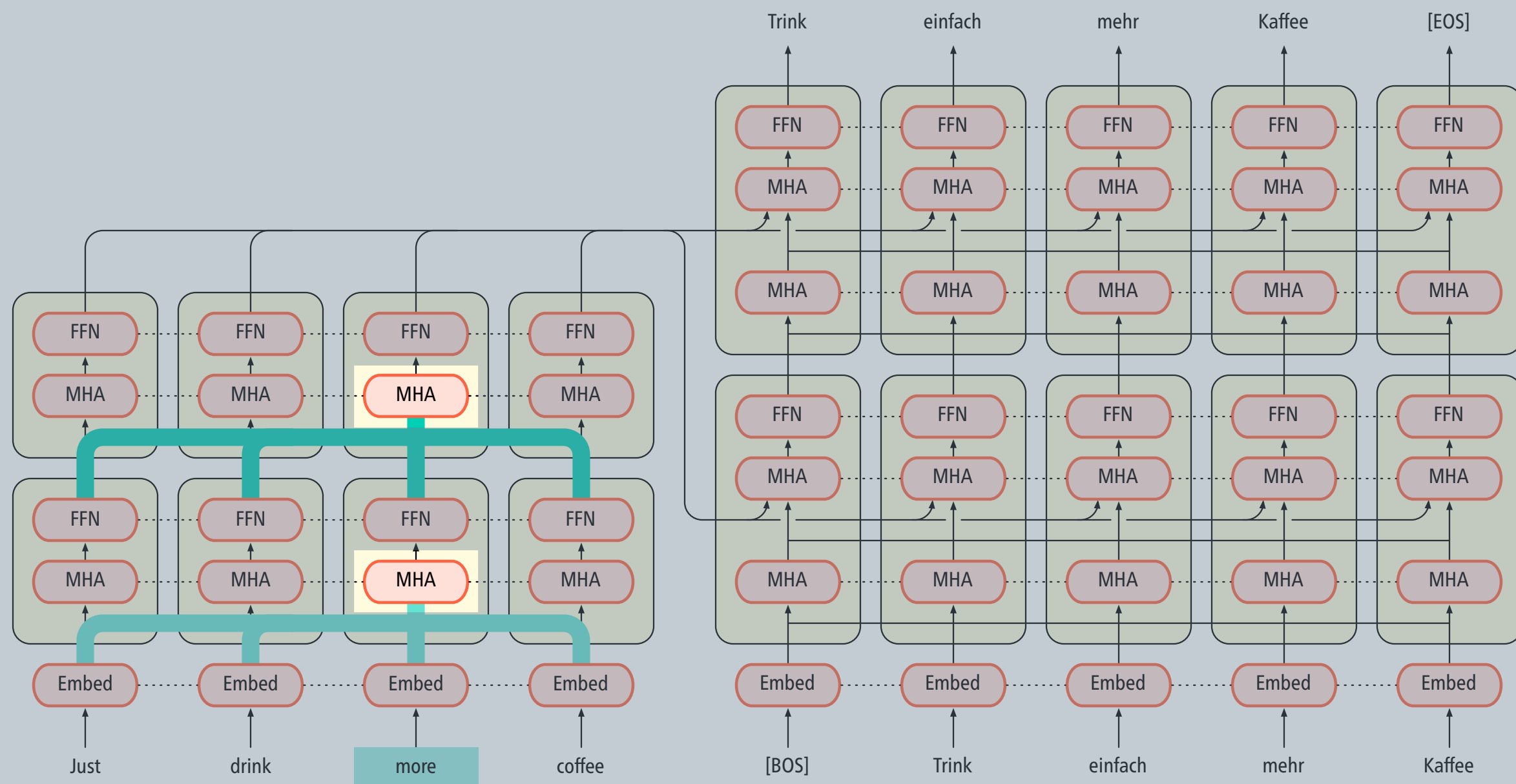


Example translation

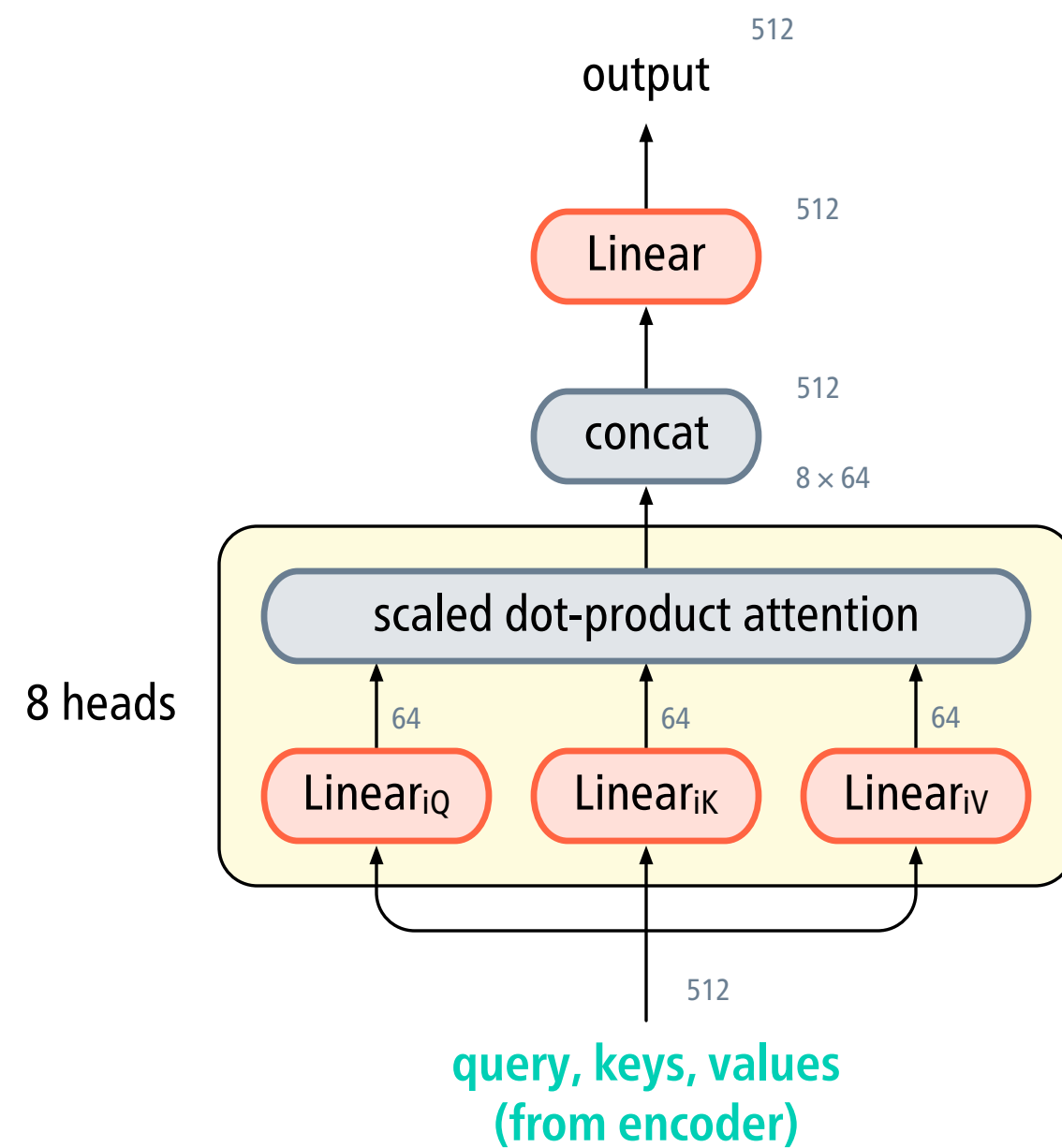


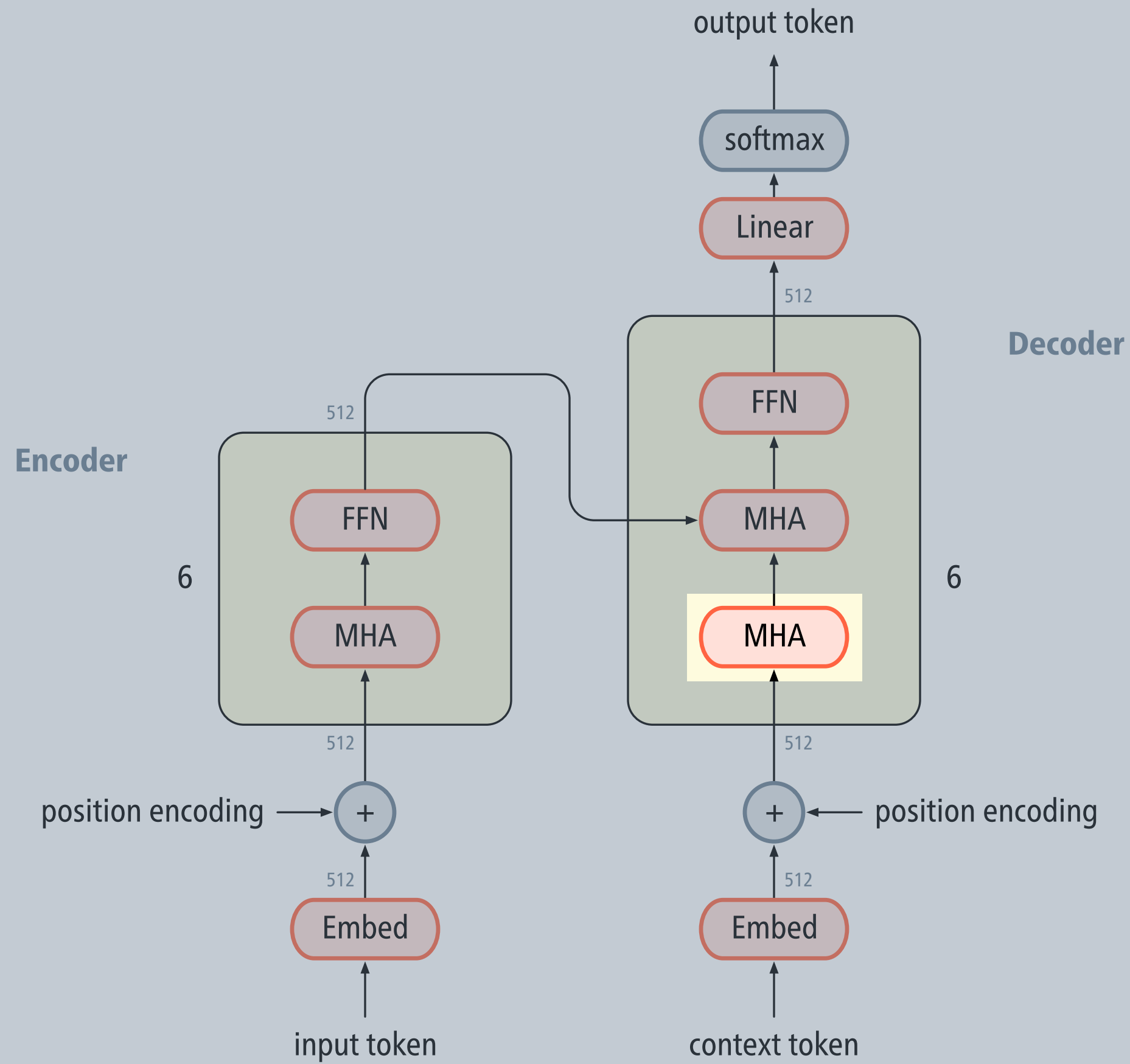


Multi-head attention in the encoder

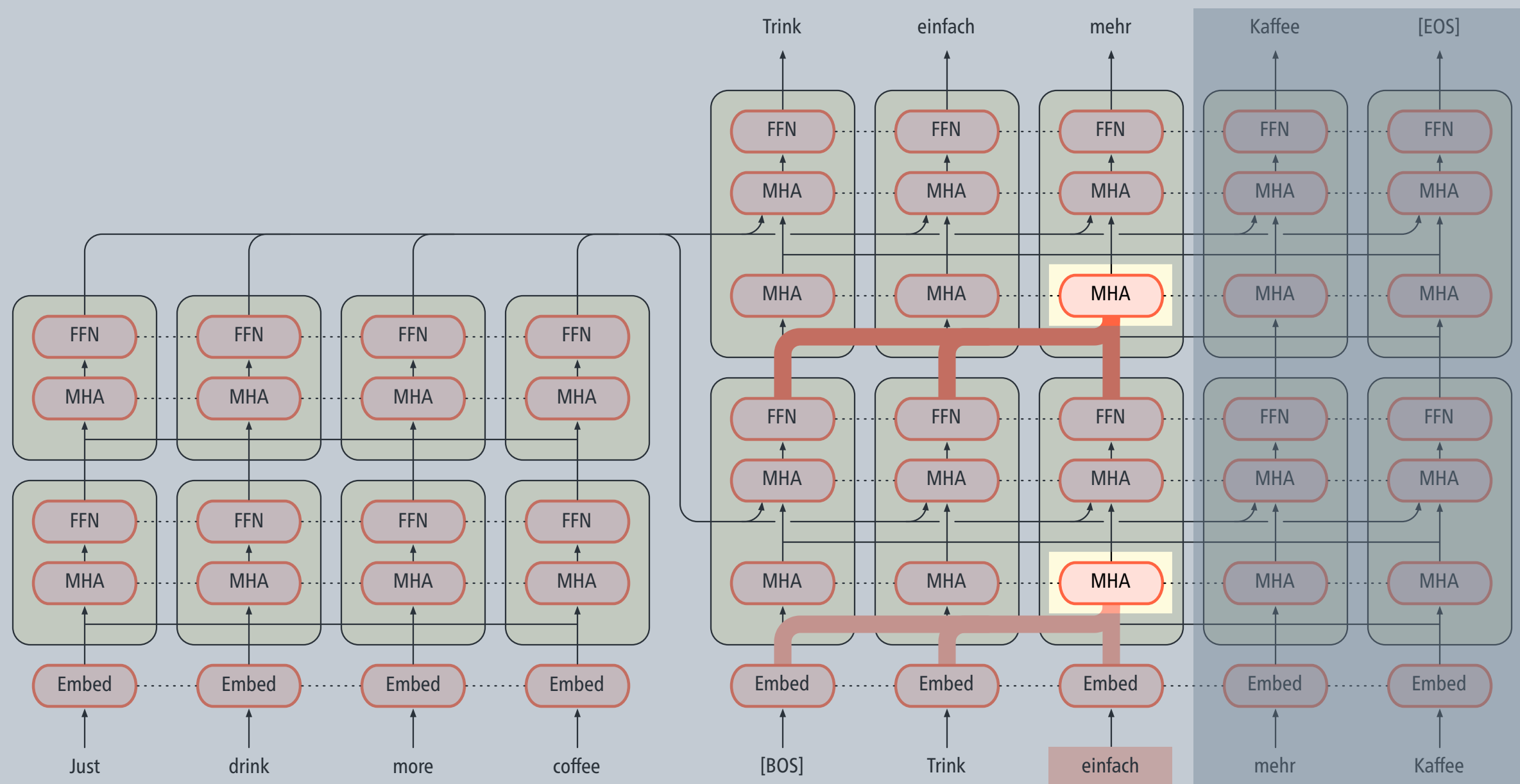


Multi-head attention in the encoder

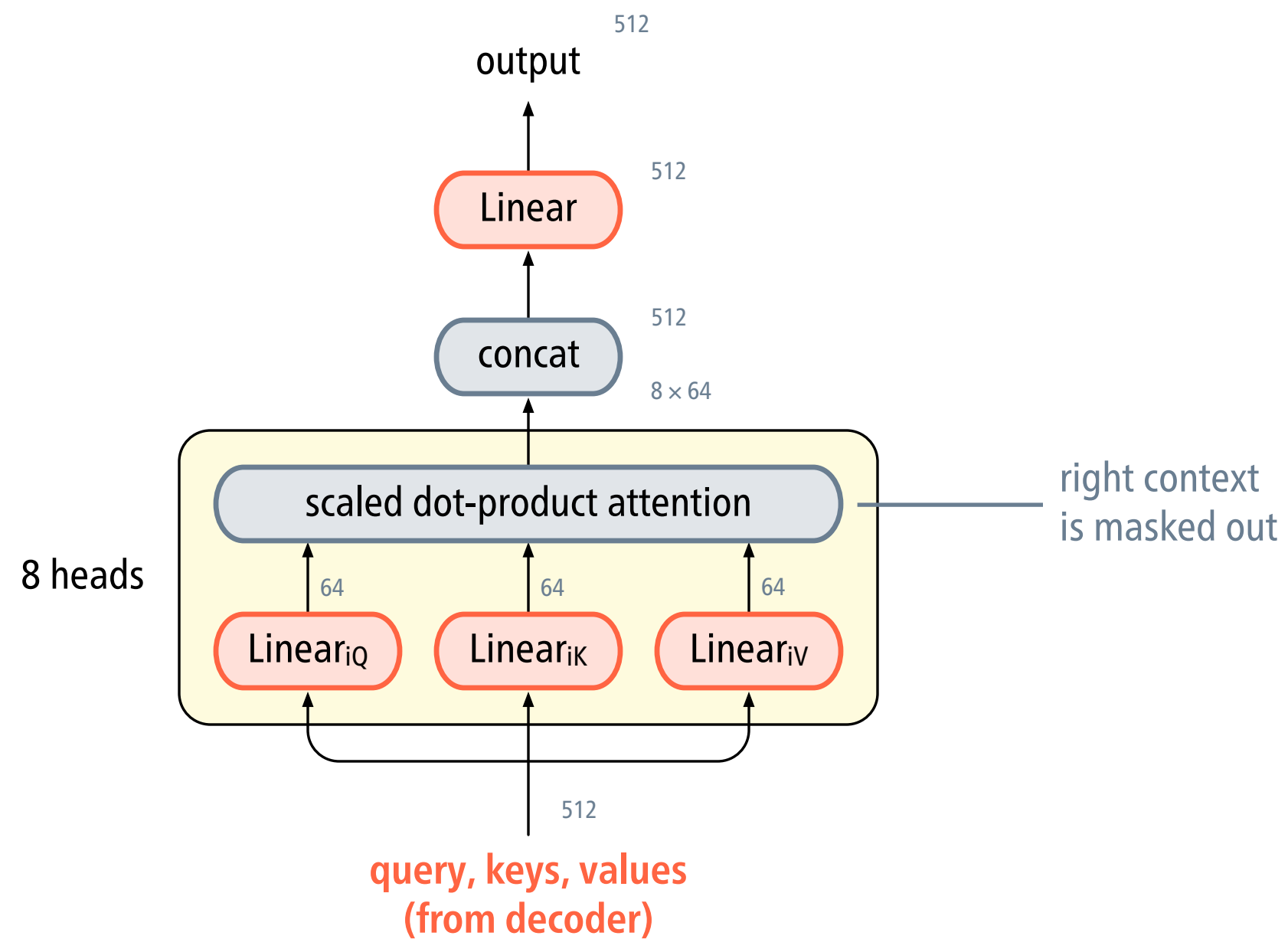


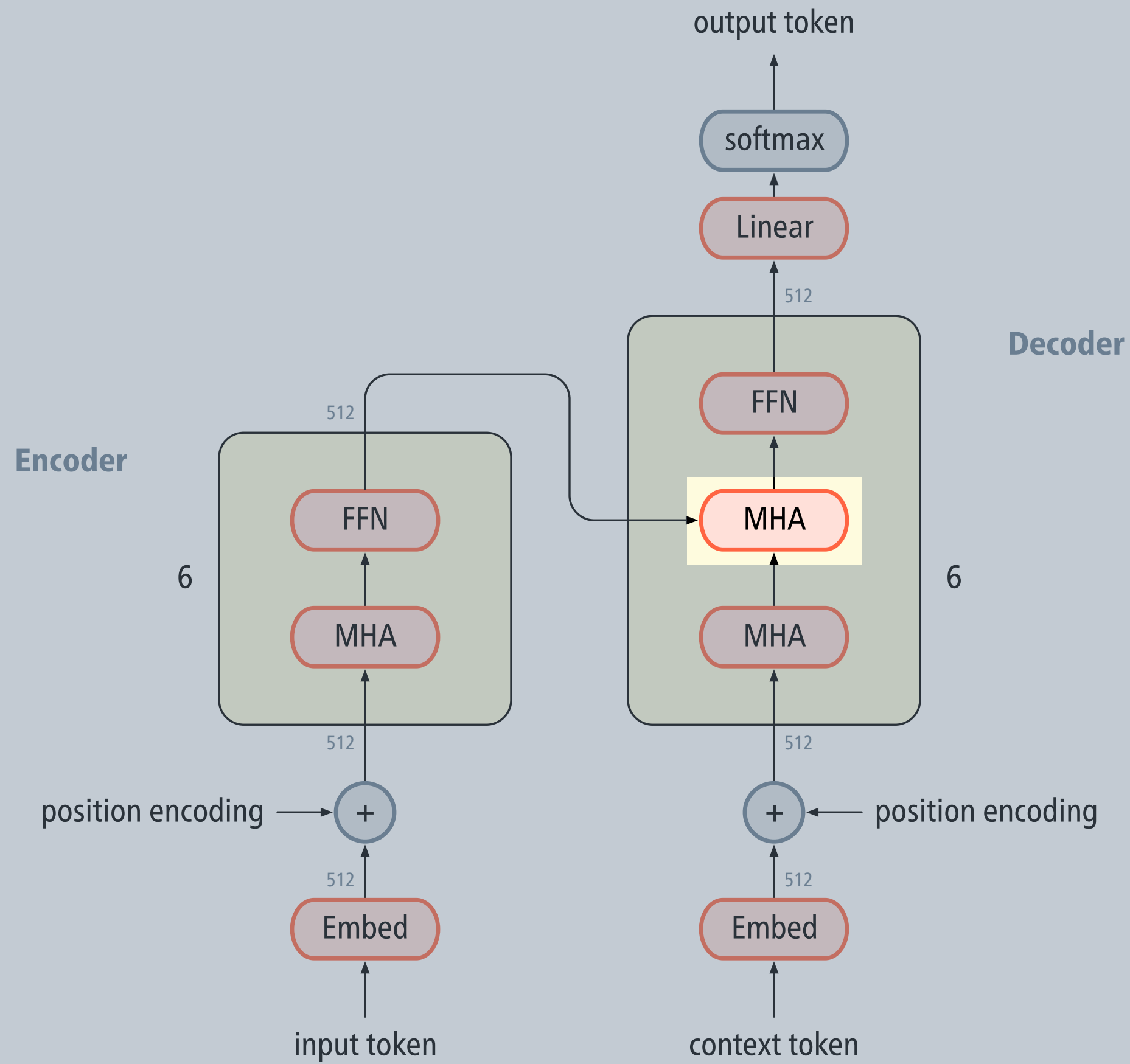


Multi-head attention in the decoder

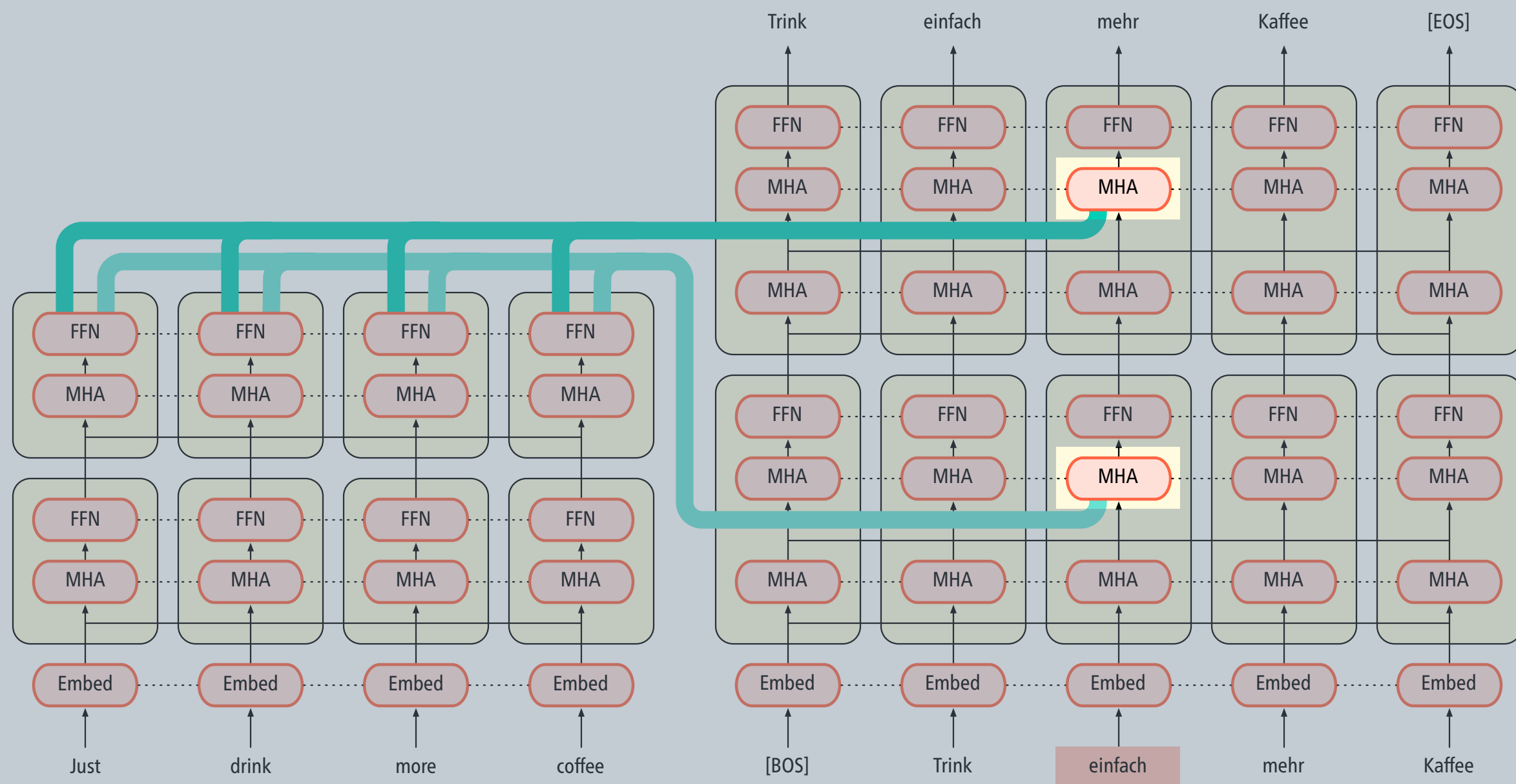


Multi-head attention in the decoder

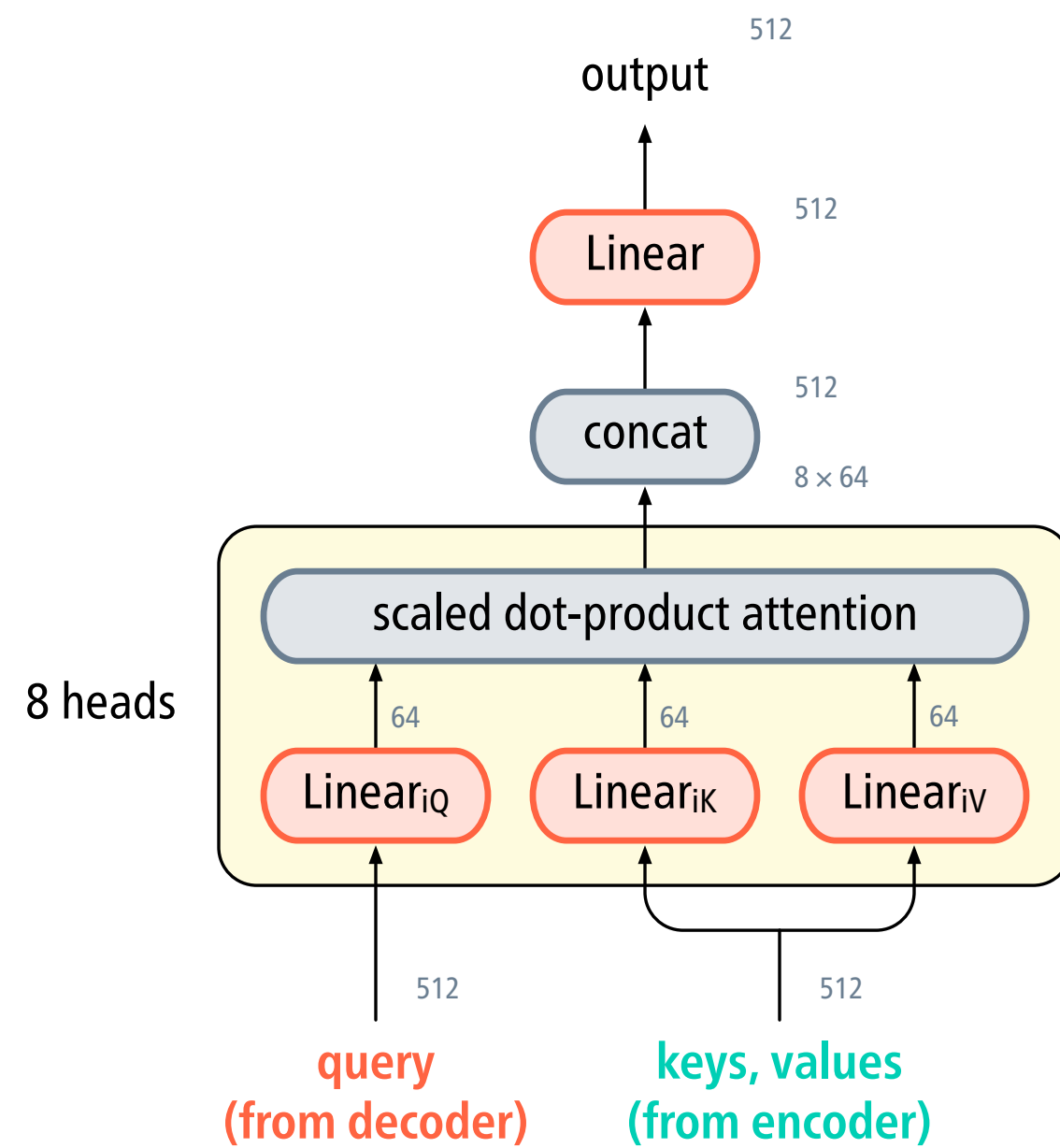


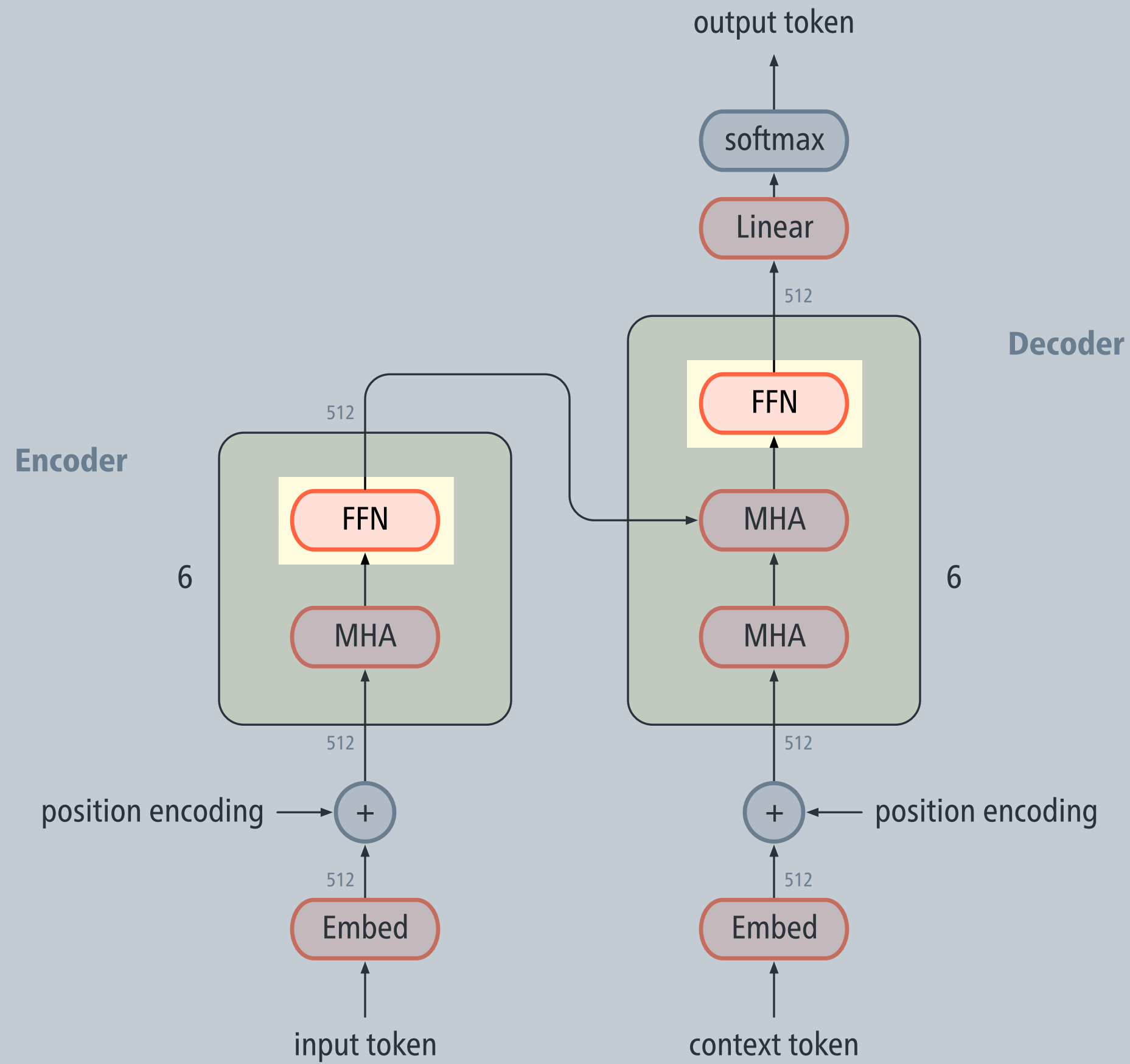


Cross-attention

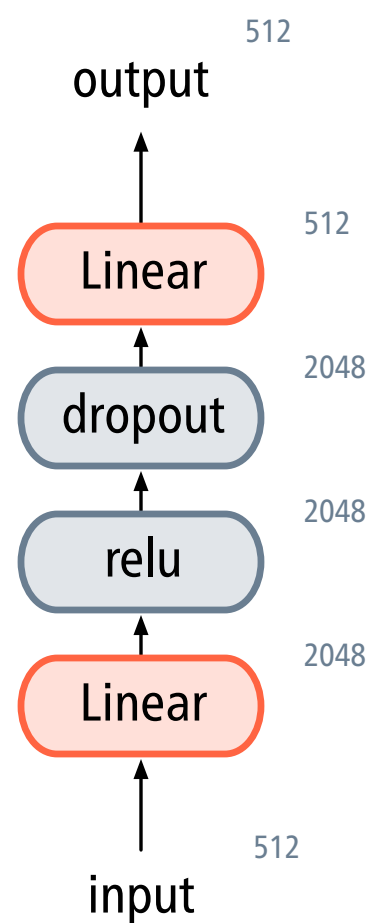


Cross-attention

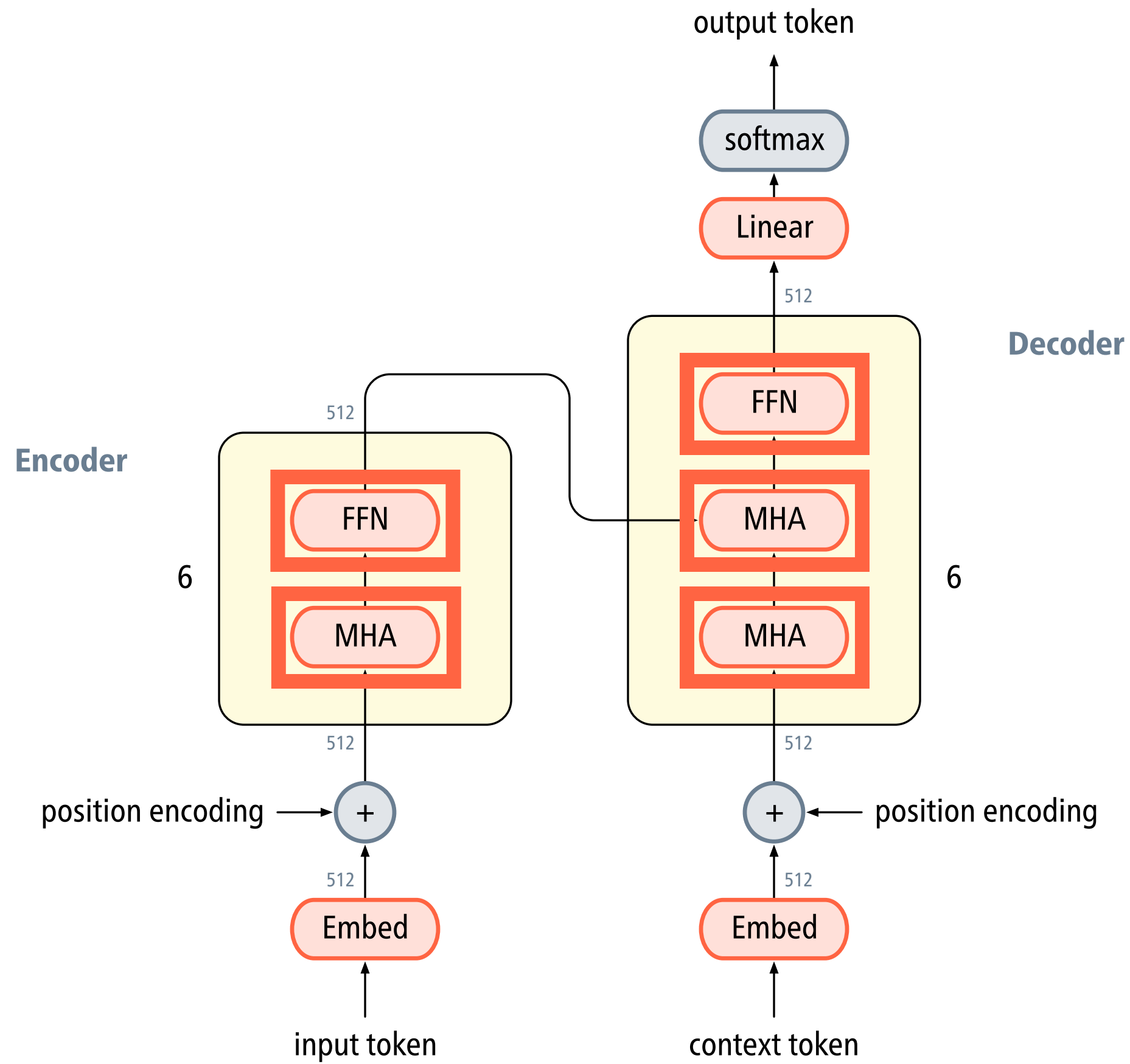




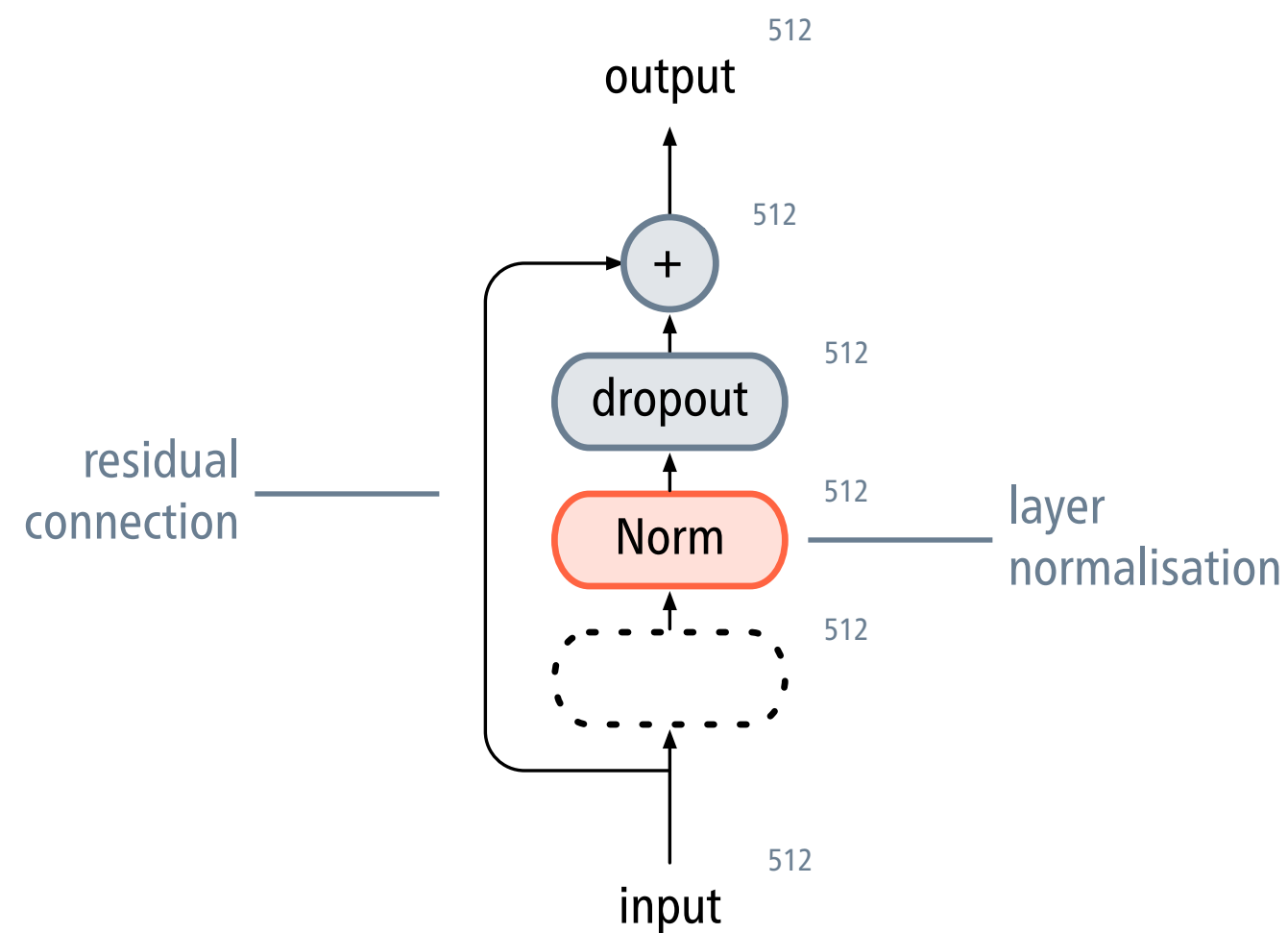
Position-wise feed-forward network



Parameters are shared across positions, but not across blocks.



Normalise-and-add wrapper



$$y = \frac{\overset{\text{gain parameter}}{\mathbf{g}}}{\sigma + \varepsilon} \odot (\mathbf{x} - \mu) + \overset{\text{bias parameter}}{\mathbf{b}}$$

$$\mu = \frac{1}{|\mathbf{x}|} \sum_{i=1}^{|\mathbf{x}|} x_i$$

$$\sigma = \sqrt{\frac{1}{|\mathbf{x}|} \sum_{i=1}^{|\mathbf{x}|} (x_i - \mu)^2}$$

Further details

- Embeddings are augmented by position encodings.
approximate encoding of absolute positions
- Training the model uses several tricks related to batching, masking, loss, and regularisation.
for details and PyTorch code, see the [‘Annotated Transformer’](#)

Translation performance

	BLEU	FLOPs
GNMT + RL (Wu et al., 2016)	39.92	$1.4 \cdot 10^{20}$
ConvS2S (Gehring et al., 2017)	40.46	$1.5 \cdot 10^{20}$
MoE (Shazeer et al., 2017)	40.56	$1.2 \cdot 10^{20}$
Transformer (big model)	41.80	$2.4 \cdot 10^{19}$

BLEU score and training cost (FLOPs) on the English-to-French newstest2014 test data | [Vaswani et al. \(2017\)](#)