Natural Language Processing

Lecture 5.4

The Transformer architecture

Marco Kuhlmann
Department of Computer and Information Science

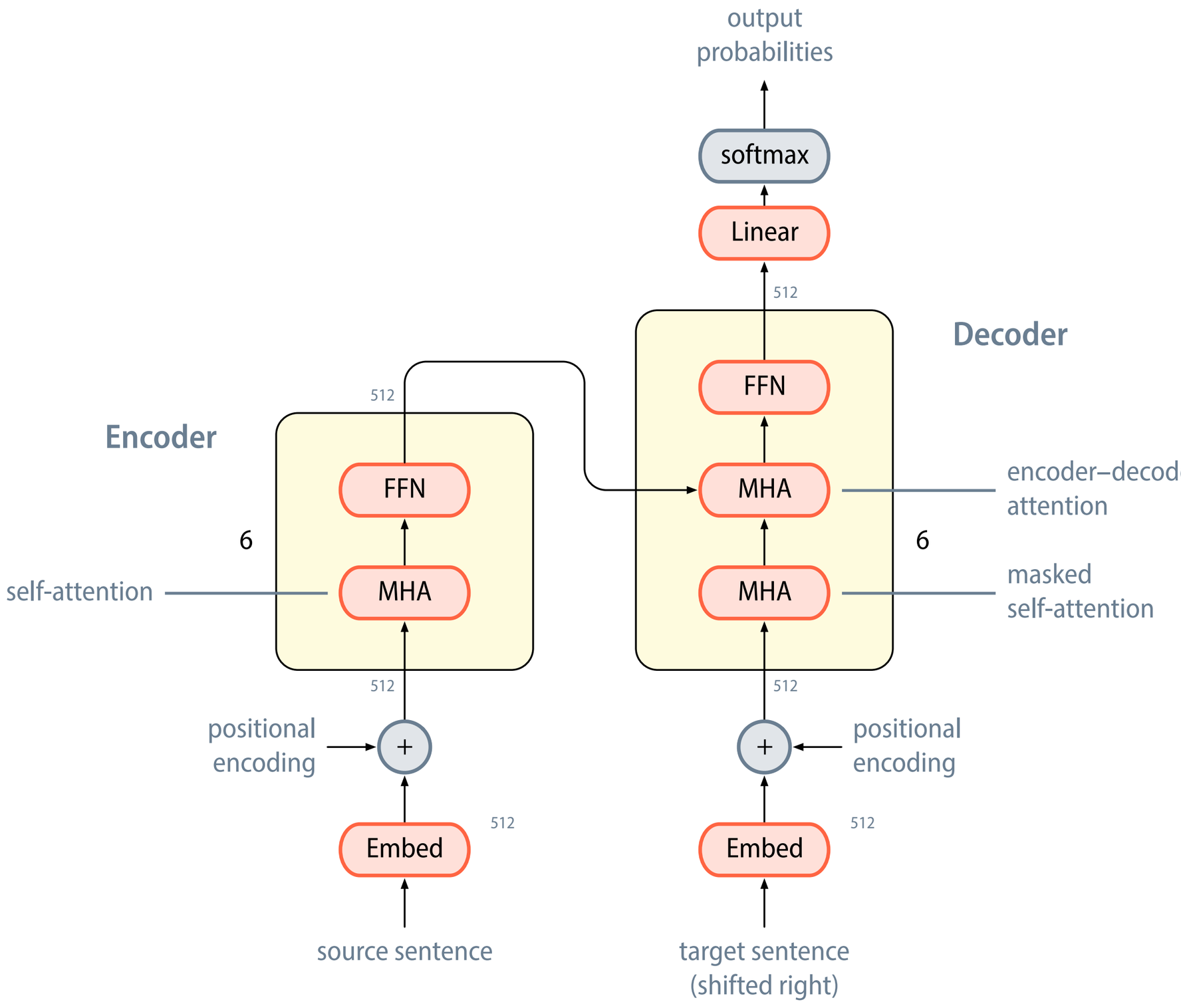
Attention is all you need

- The attention mechanism allows the direct modelling of dependencies between words, without regard to their distance.
- However, recurrent neural networks implement a mode of computation that is essentially sequential.

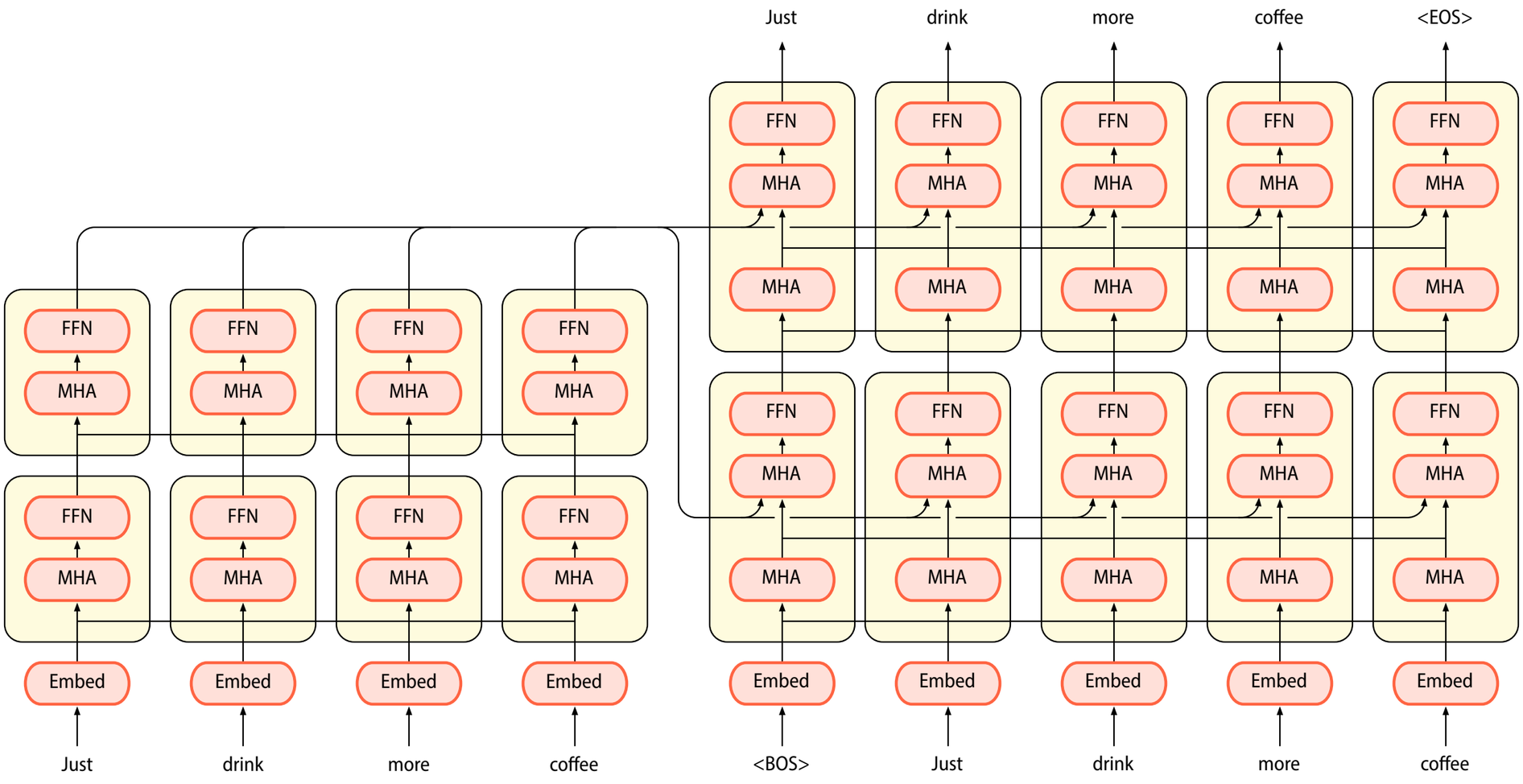
  precludes parallelisation

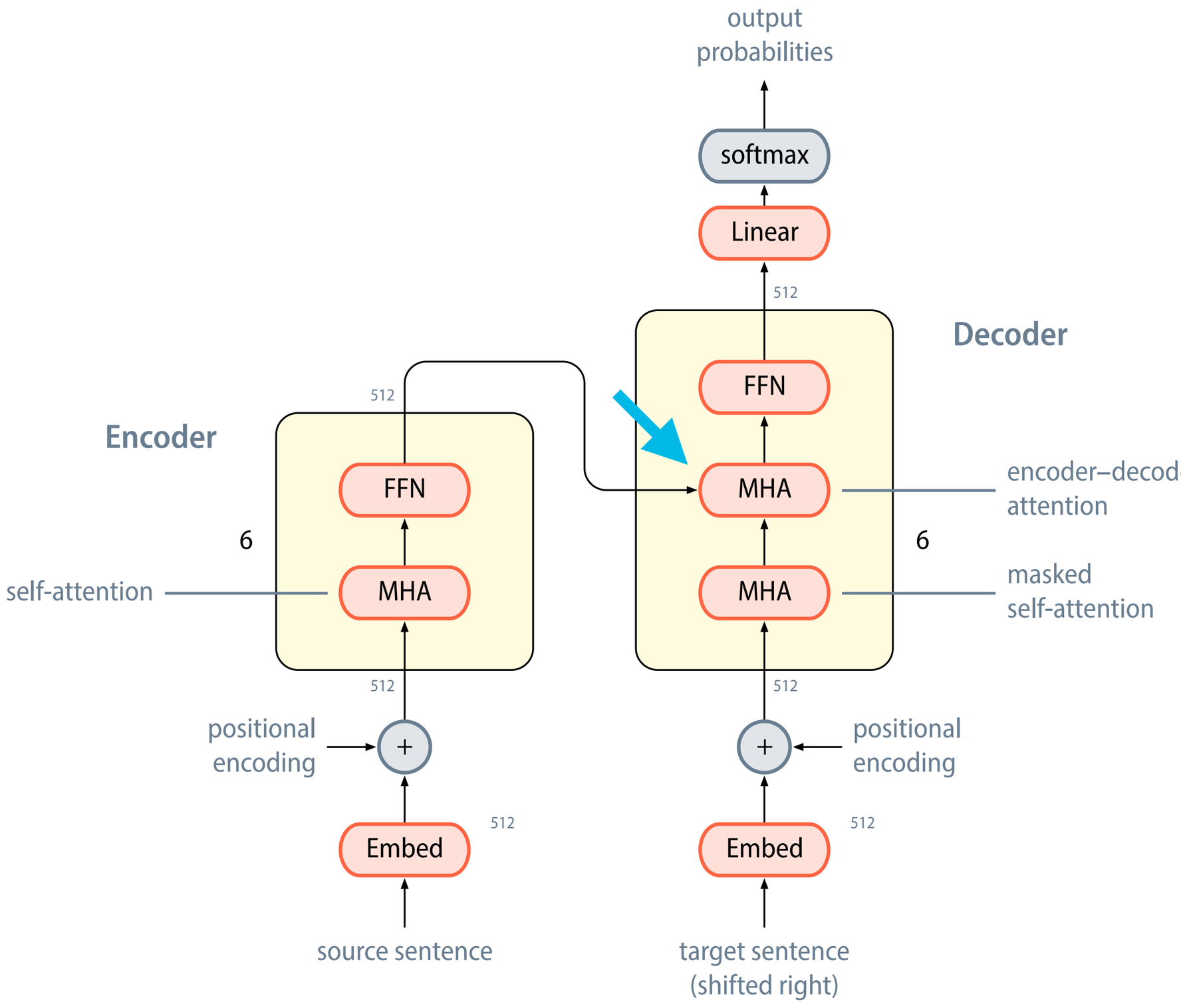
- The **Transformer** is an architecture that eschews recurrence and instead relies entirely on an attention mechanism.

Vaswani et al. (2017)

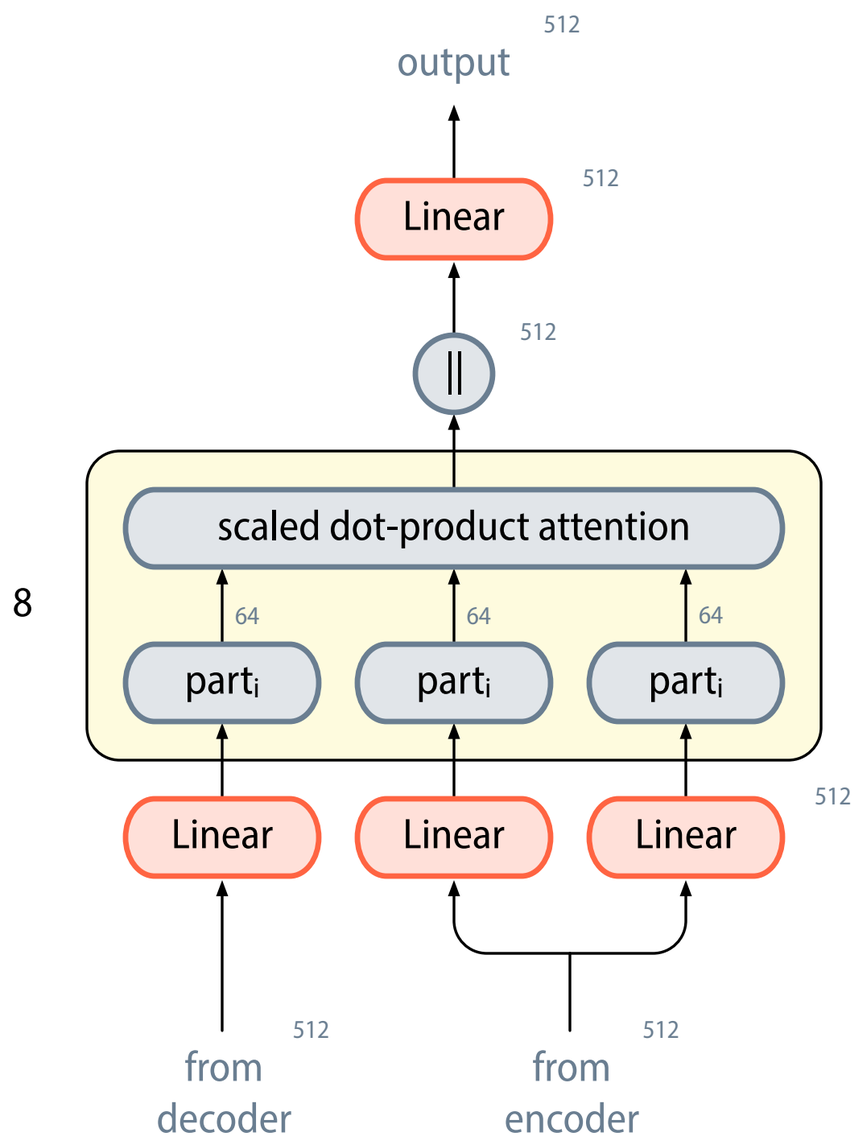

output
probabilities
softmax
Linear
512
Decoder
FFN
MHA
encoder–decoder
attention
masked
self-attention
6
MHA
FFN
512
Encoder
6
self-attention
MHA
512
positional
encoding
positional
encoding
512
Embed
Embed
512
source sentence
target sentence
(shifted right)

Example translation

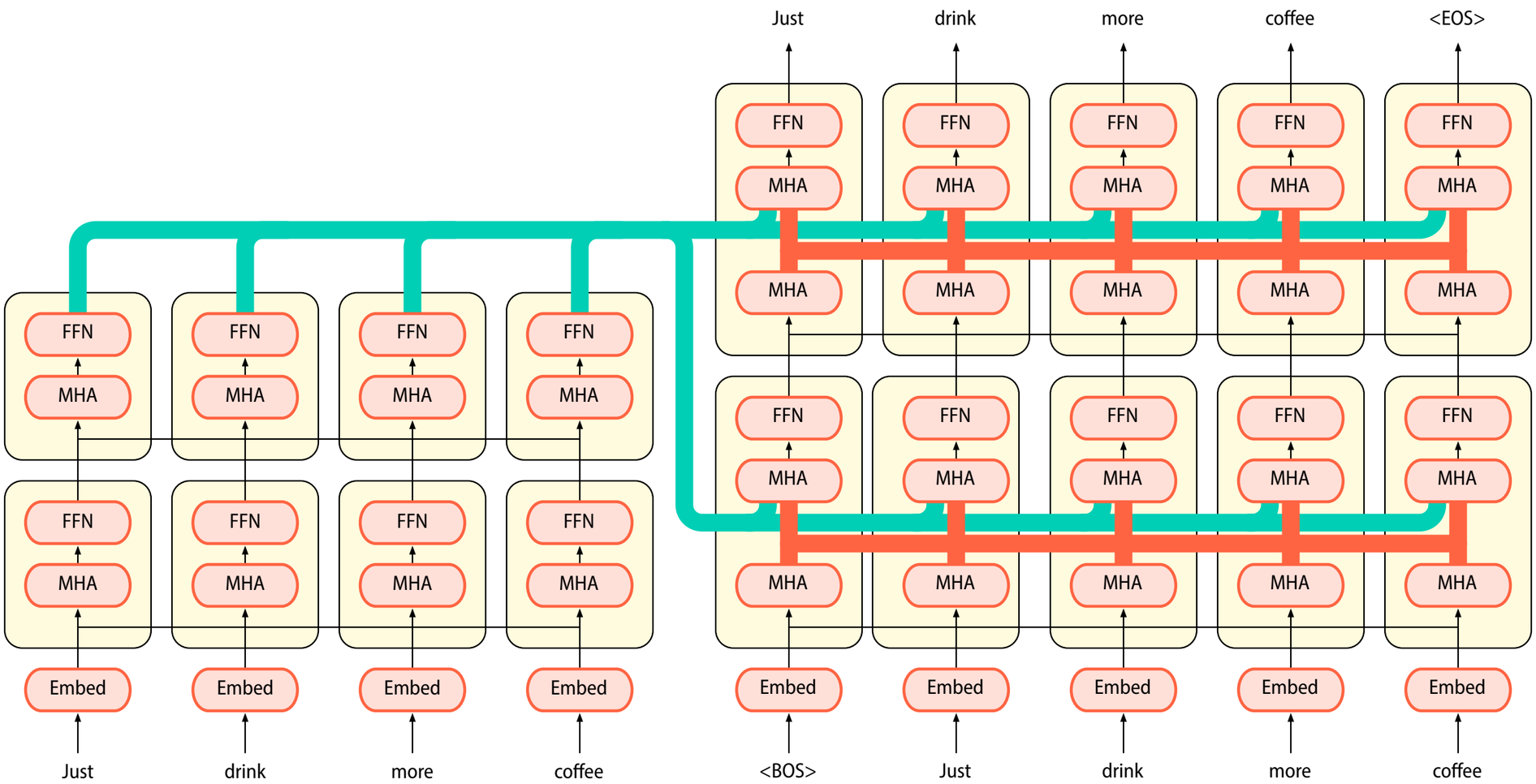


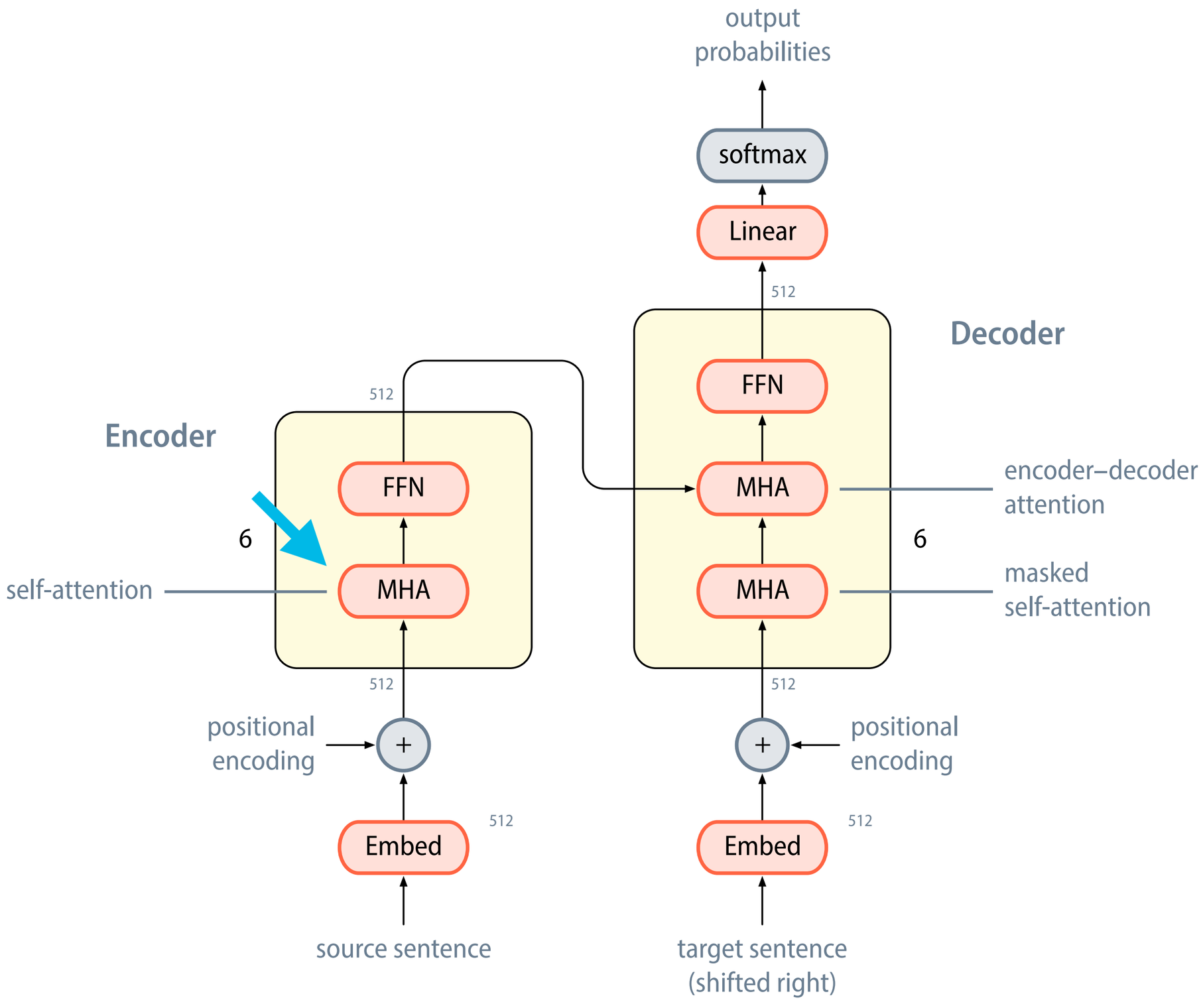

output
probabilities
softmax
Linear
512
Decoder
FFN
MHA
encoder–decoder
attention
6
masked
self-attention
MHA
512
positional
encoding
+
512
Embed
target sentence
(shifted right)
Encoder
512
FFN
6
self-attention
MHA
512
positional
encoding
+
512
Embed
source sentence

Encoder–decoder attention

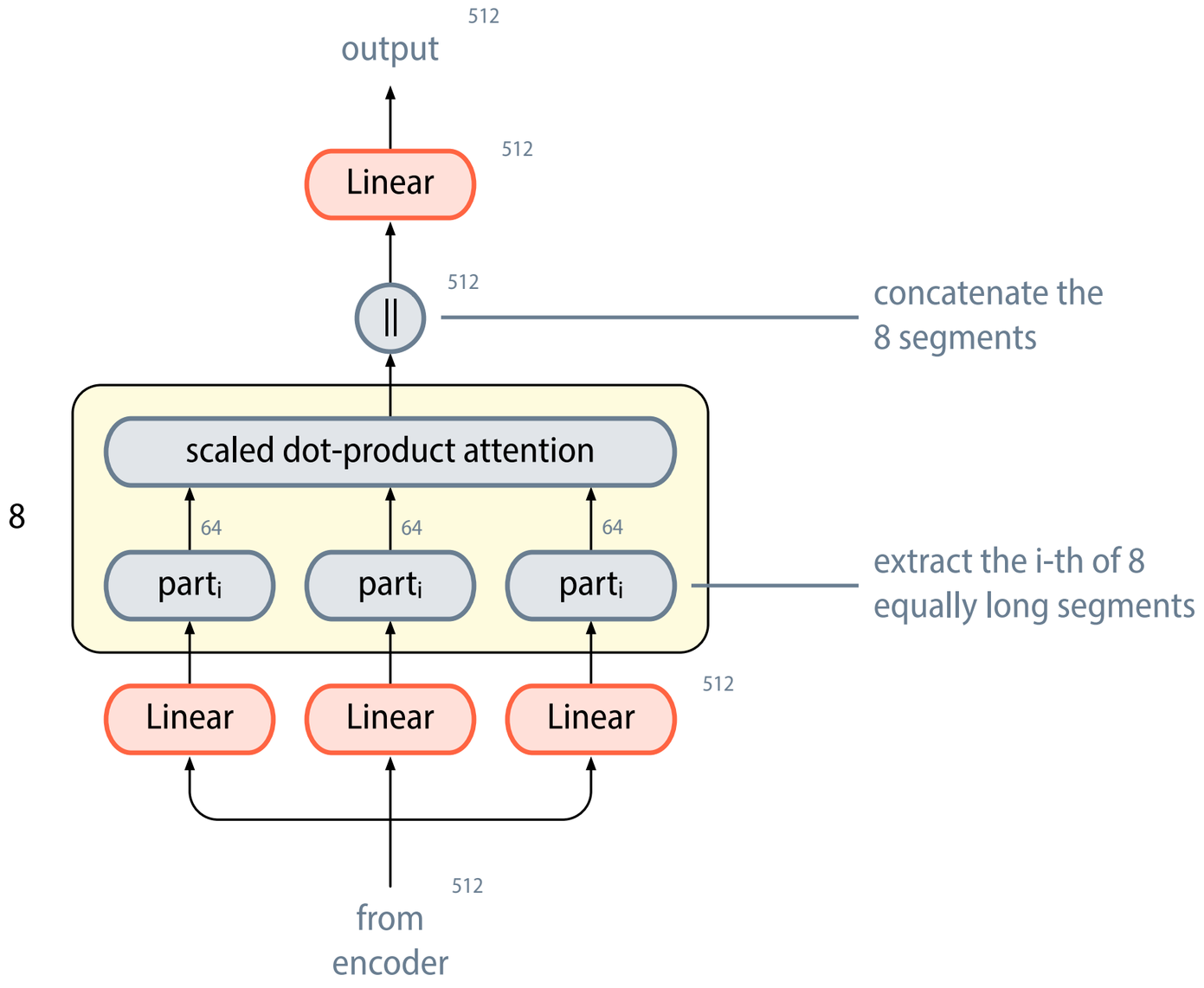


Encoder–decoder attention

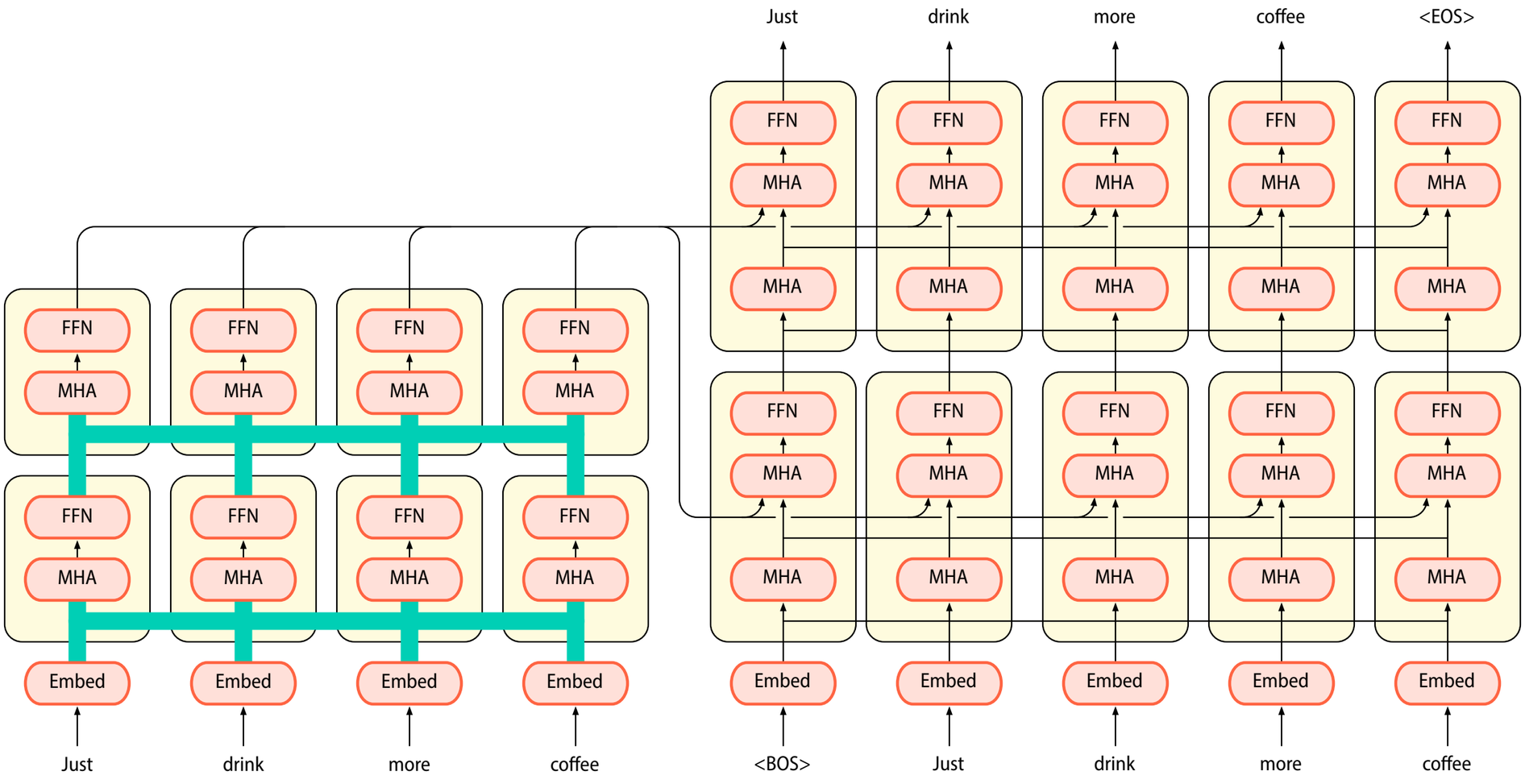


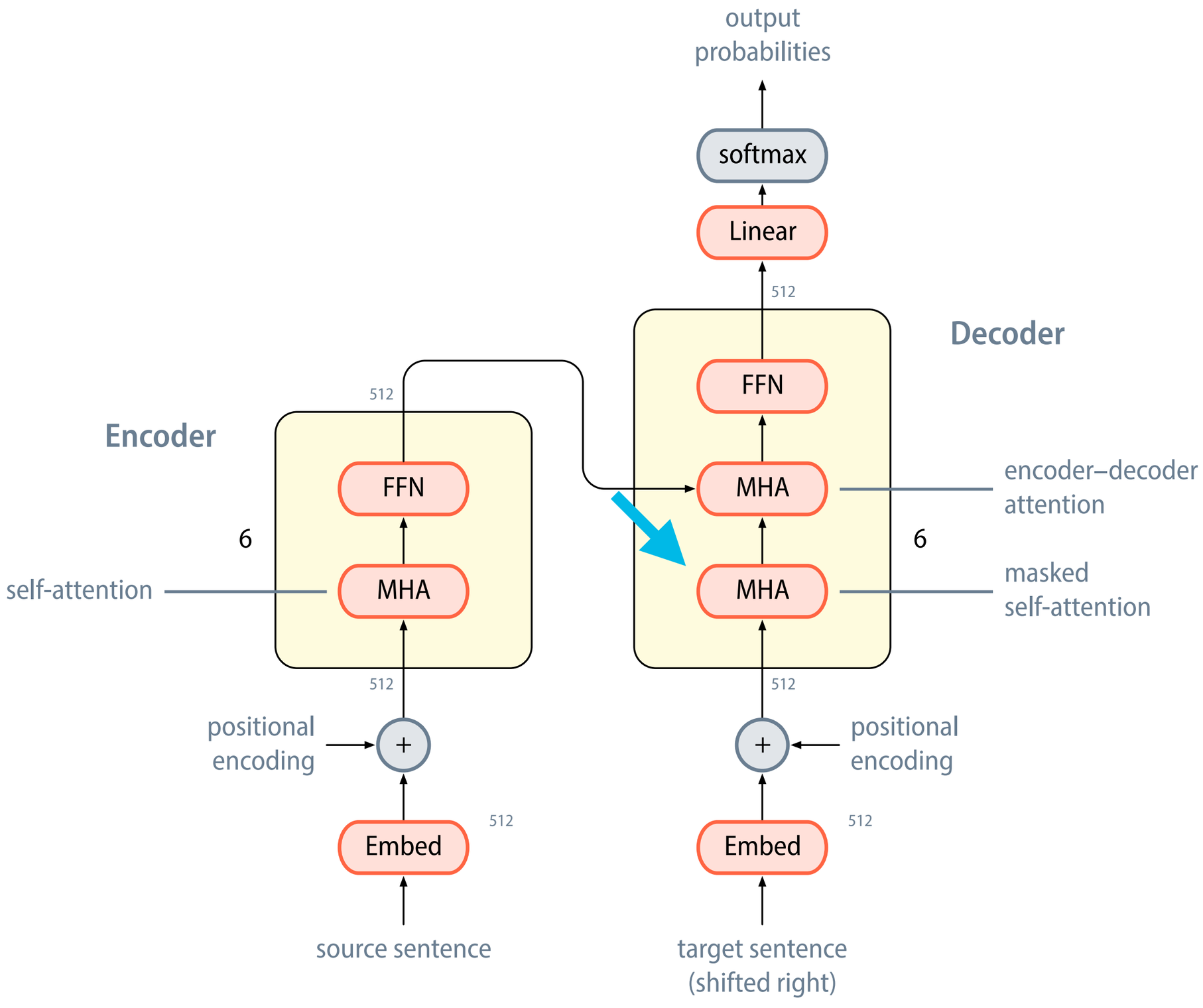

output
probabilities
softmax
Linear
512
Decoder
FFN
MHA
encoder–decoder
attention
6
masked
self-attention
MHA
512
positional
encoding
512
Embed
target sentence
(shifted right)
Encoder
512
FFN
6
self-attention
MHA
512
positional
encoding
512
Embed
source sentence

Self-attention (encoder)

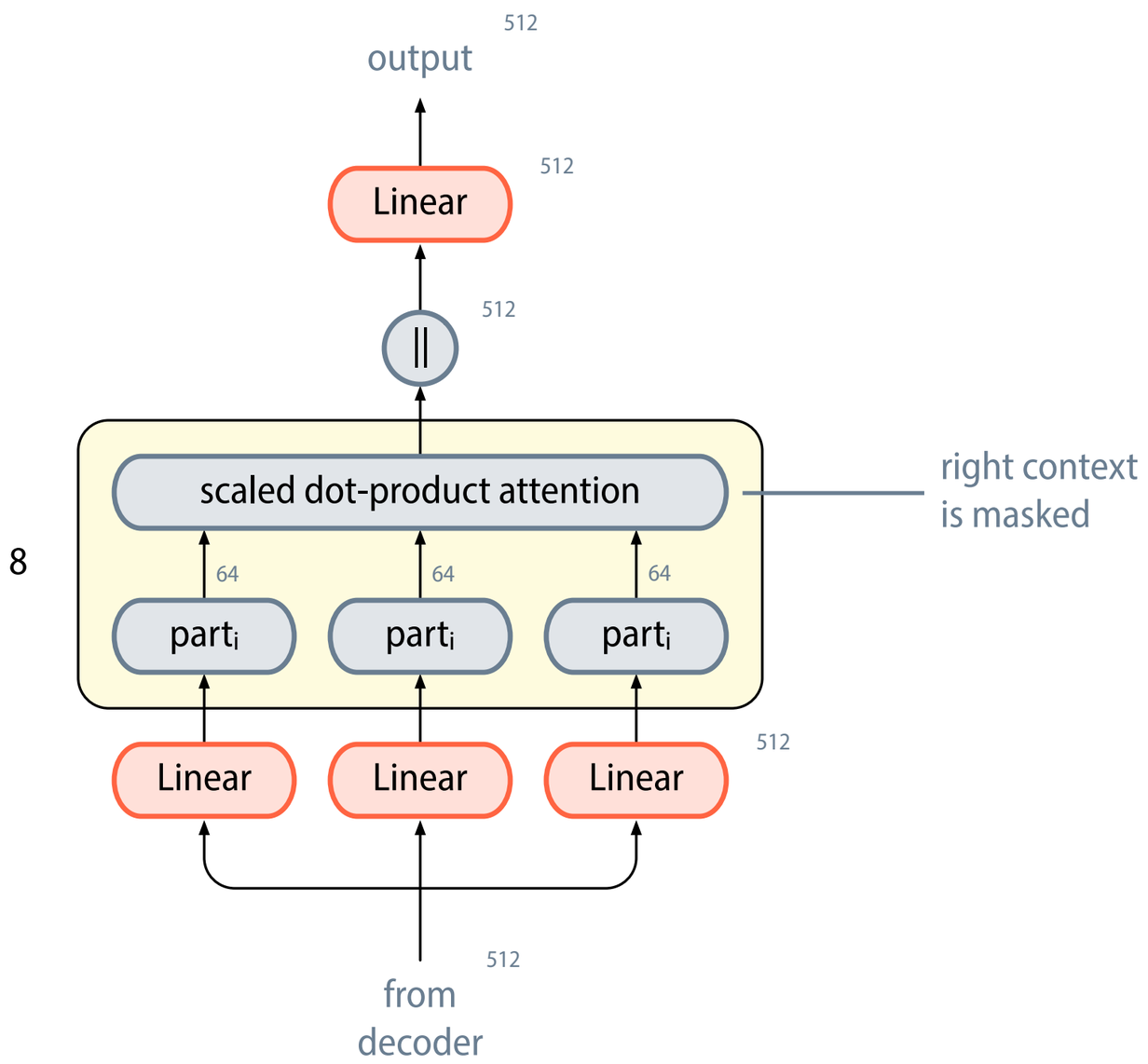


Self-attention (encoder)

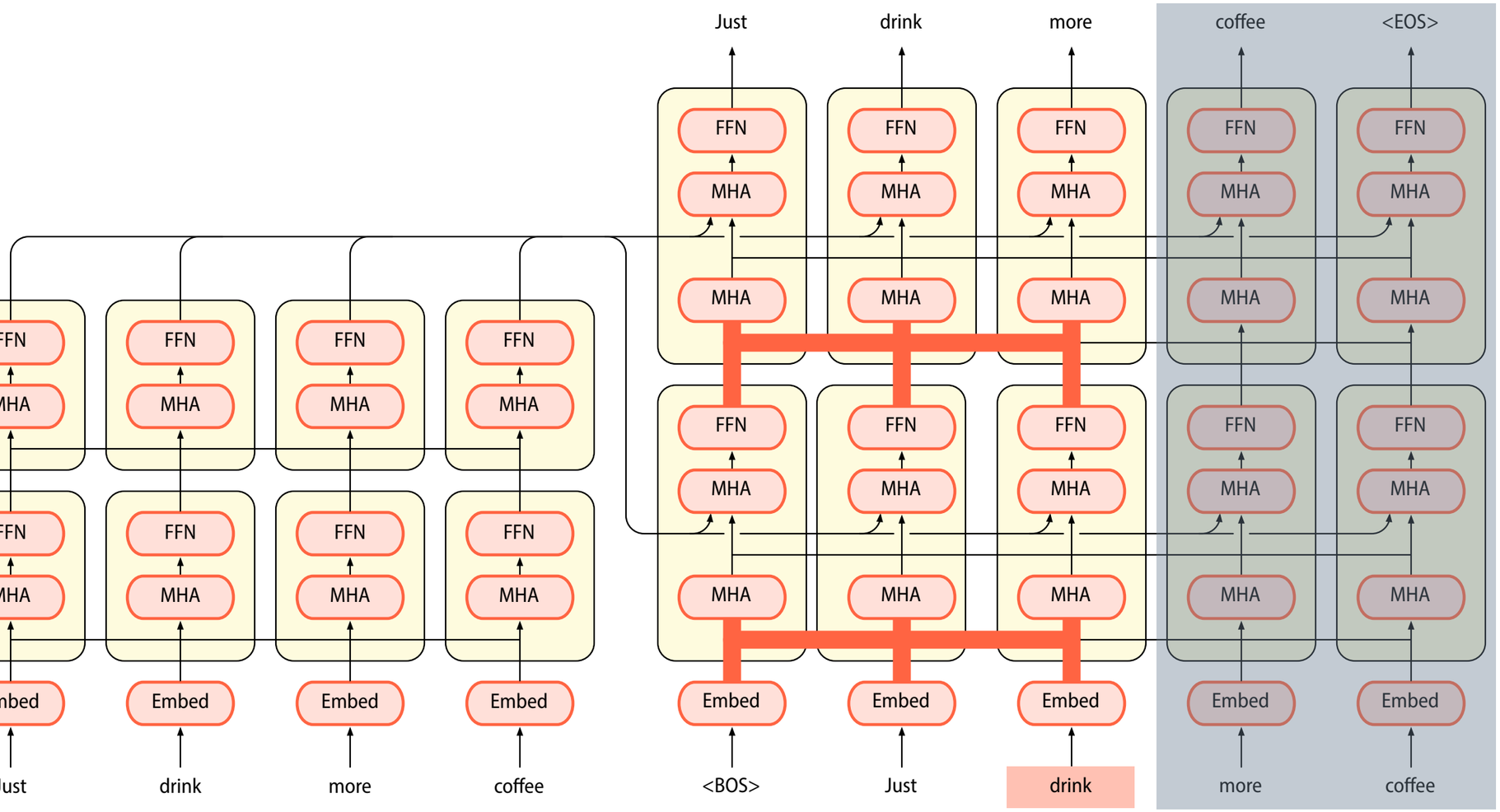



output
probabilities
softmax
Linear
512
Decoder
FFN
MHA
encoder–decoder
attention
6
masked
self-attention
MHA
512
Encoder
512
FFN
6
self-attention
MHA
512
positional
encoding
positional
encoding
512
Embed
512
Embed
source sentence
target sentence
(shifted right)

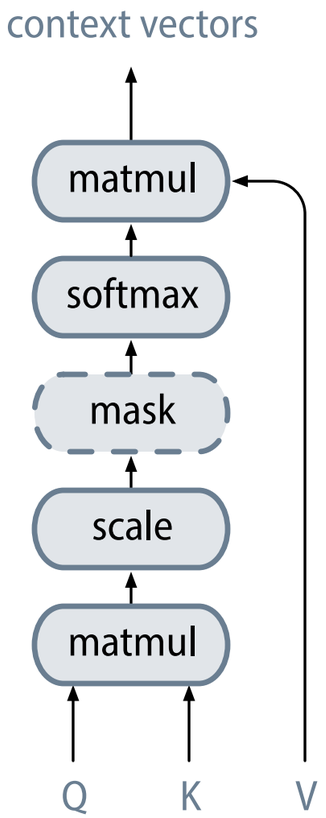
Masked self-attention (decoder)



Masked self-attention (decoder)



Scaled dot-product attention

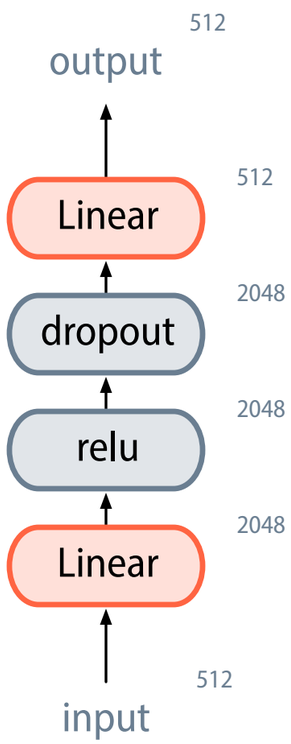


- The input consists of queries and keys of dimension d_K, and values of dimension d_V.
- The context vectors are computed as

$$\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_K}}\right)\boldsymbol{V}$$

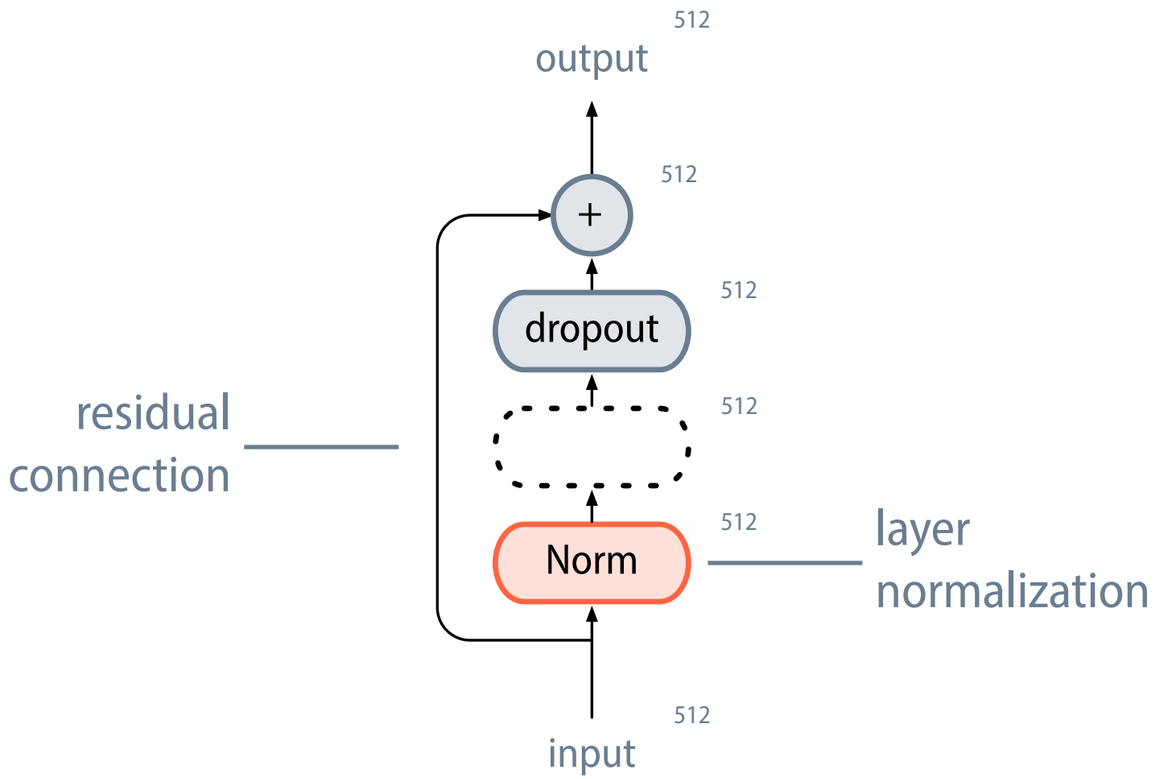
- Scaling prevents the softmax from being pushed into regions with small gradients.

Position-wise feed-forward network



Parameters are shared across positions, but not across blocks.

Normalise-and-add wrapper



gain parameter

bias parameter

$$\boldsymbol{y} = \frac{\boldsymbol{g}}{\sigma + \varepsilon} \odot (\boldsymbol{x} - \mu) + \boldsymbol{b}$$

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

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

Further details

- Embeddings are defined on word pieces, and the weights of the embedding layers are shared between encoder and decoder.
- Embeddings are augmented by positional encodings.

  approximate encoding of absolute/relative position

- Training the model uses a number of training tricks related to batching, masking, loss, and regularisation.

  for details and PyTorch code, see the 'Annotated Transformer'

Performance

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

BLEU score and training cost (FLOPs) on the English-to-French newstest2014 test data