A Brief Introduction to the **L**arge **L**anguage **M**odel

Longxiao

27/02/2025

Basics of Transformer models

Attention is All You Need

Origin of LLM: NMT task

Neural Machine Translation

We want to calculate or infer:

$$P(y_i | input, y_1, y_2, \ldots, y_{i-1})$$

Something important:

Words: Embedding, Encoder-decoder structure

Sequence: RNN -> position embedding

Relation: Attention: cross -> self -> multi-head

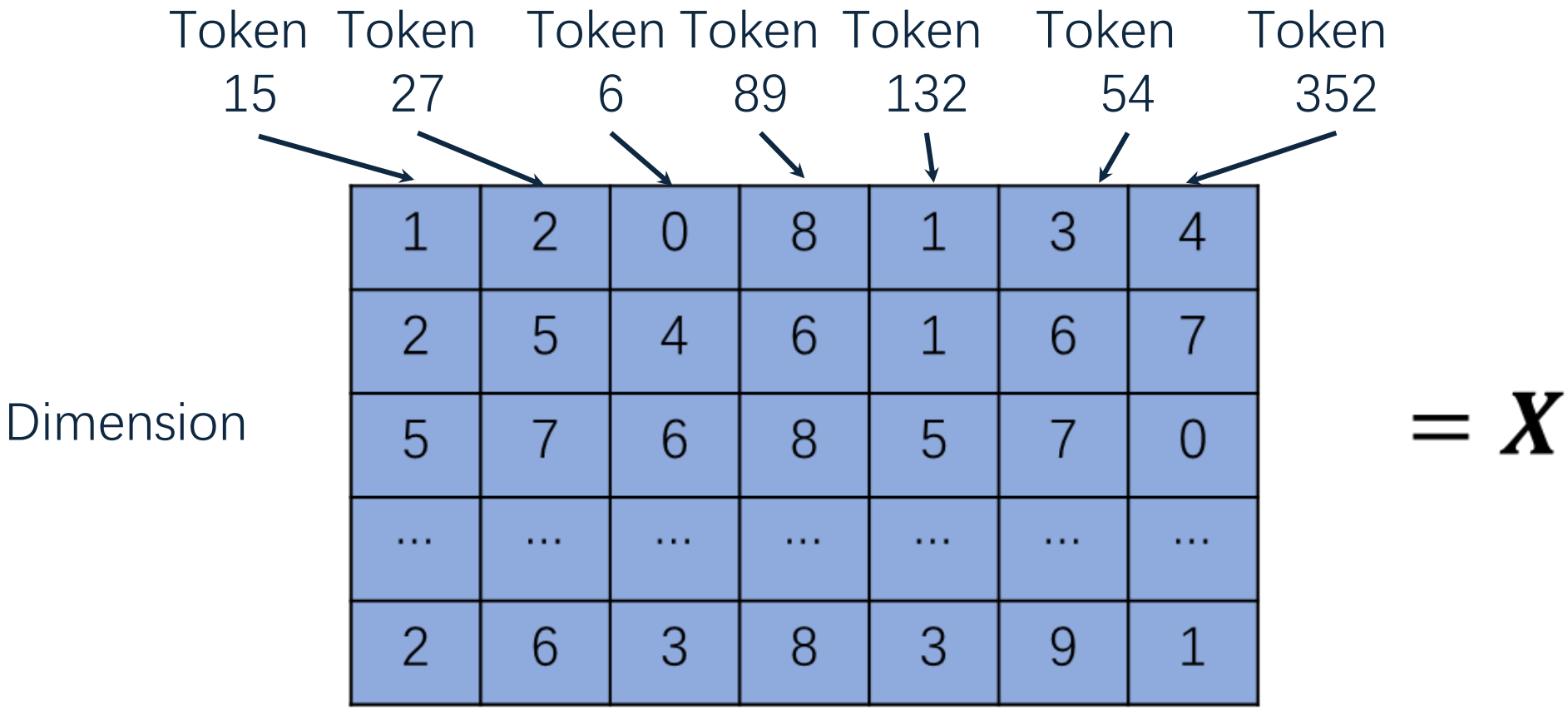
<SOS> Wir machen einen Kurs in maschinellem Lernen <EOS>

<SOS> We take a machine learning course <EOS>

Embedding

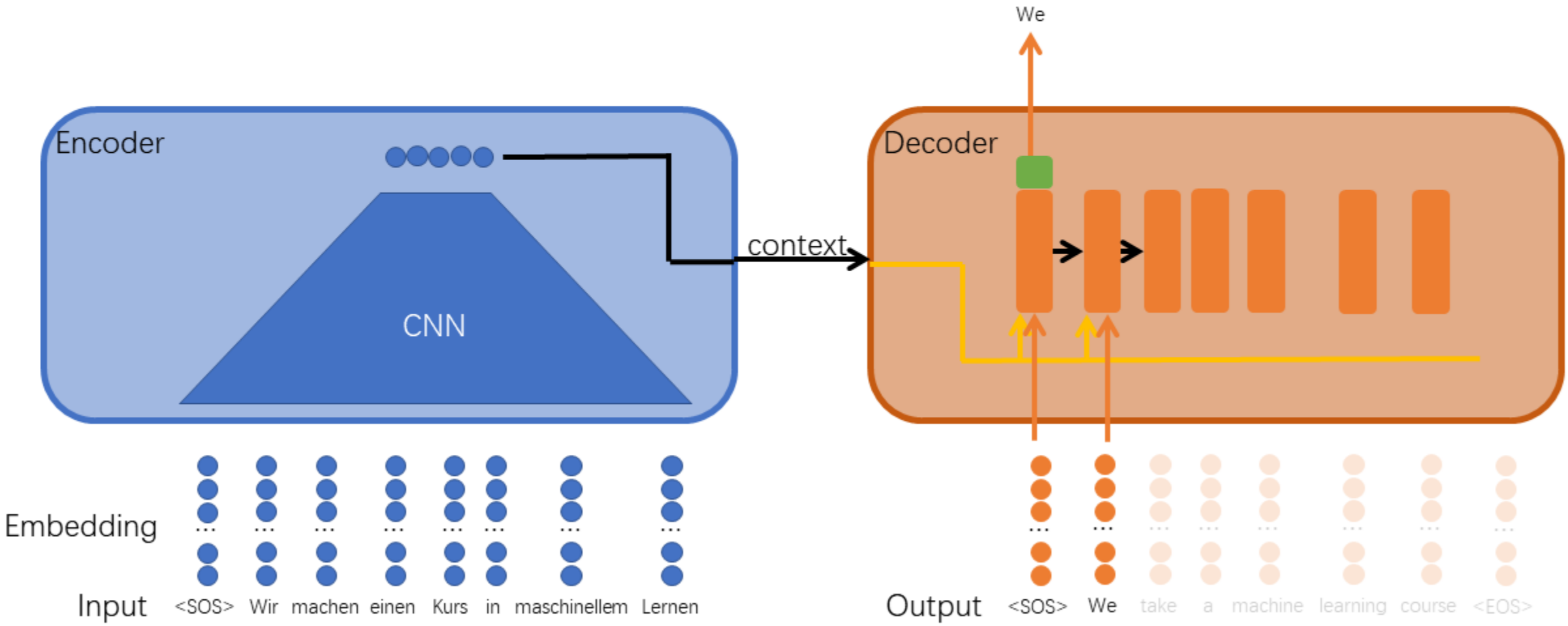
Embedding: from word to token then to vector

<SOS> Wir machen einen Kurs in maschinellem Lernen <EOS>



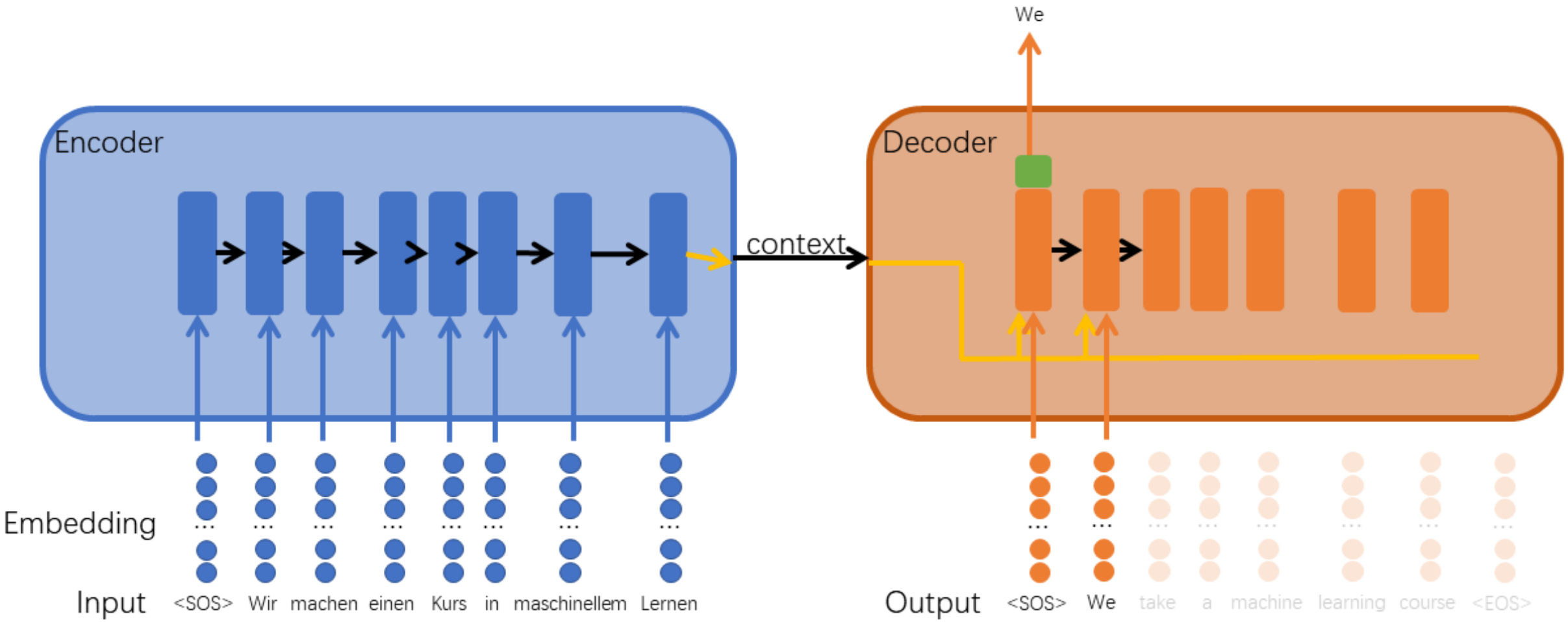
Encoder-decoder

Kalchbrenner, N., & Blunsom, P. (2013, October). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1700-1709).



Encoder-decoder

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.



Attention! please

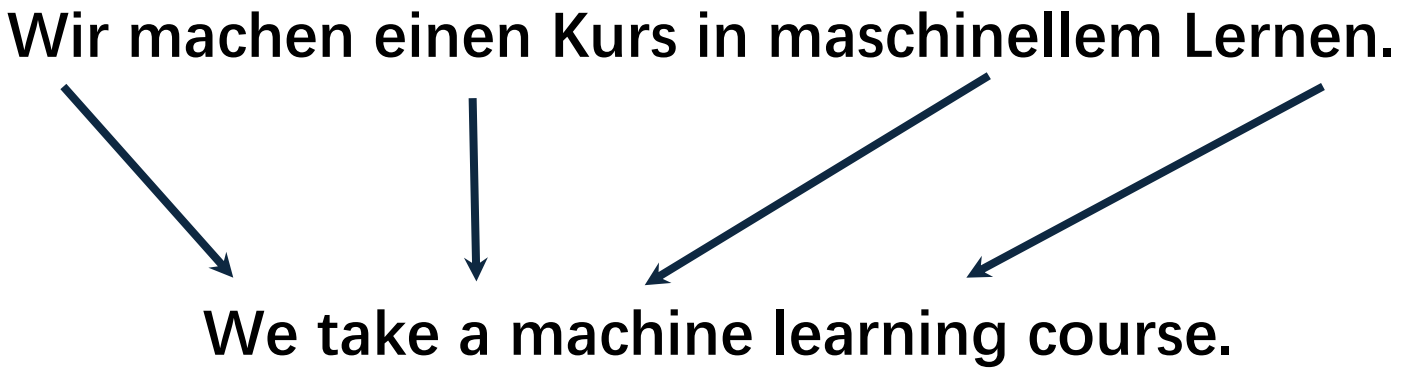
- Cross attention:

Bahdanau Attention

Luong Attention

- Self Attention

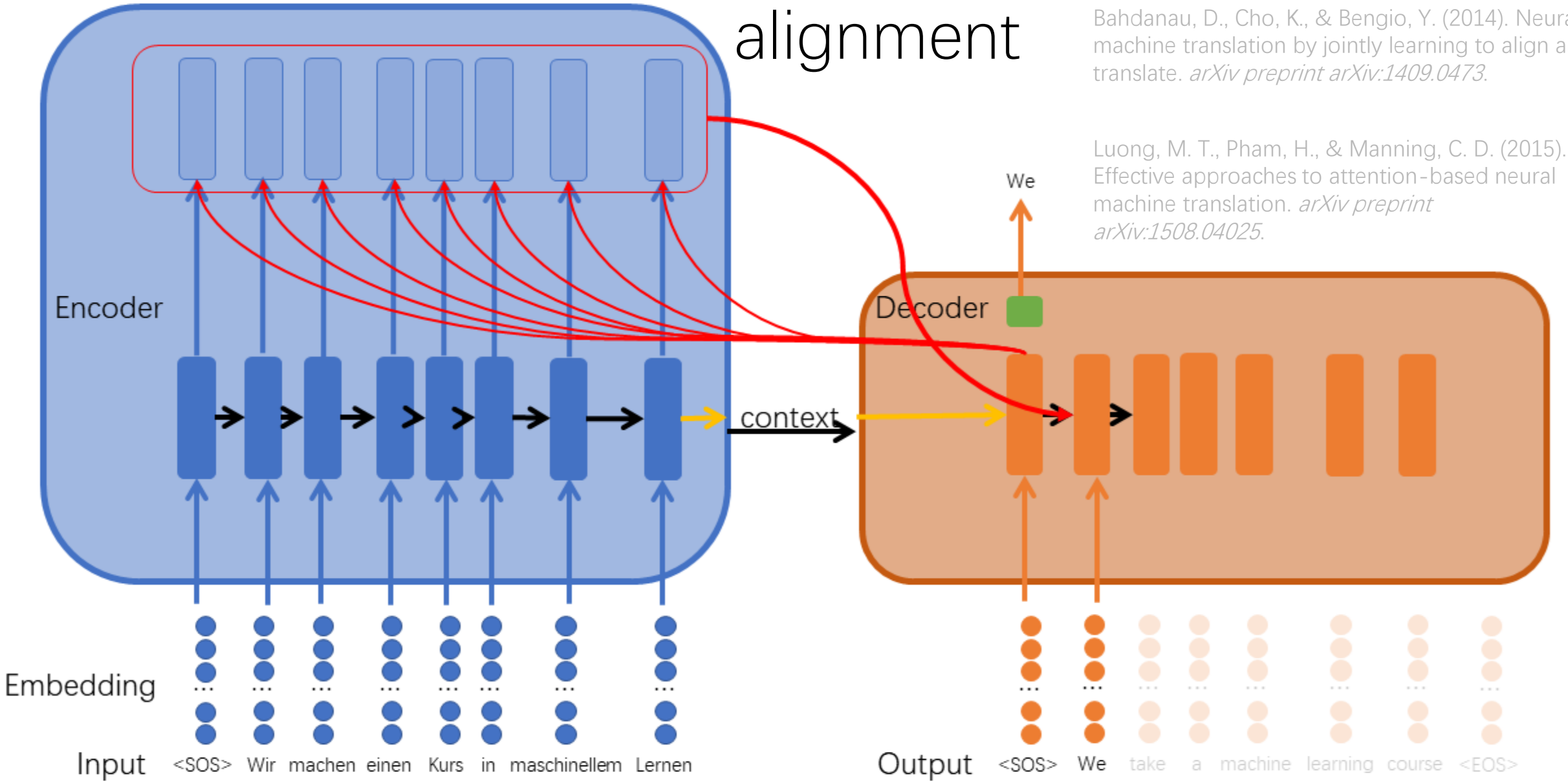
- Multi-head Attention



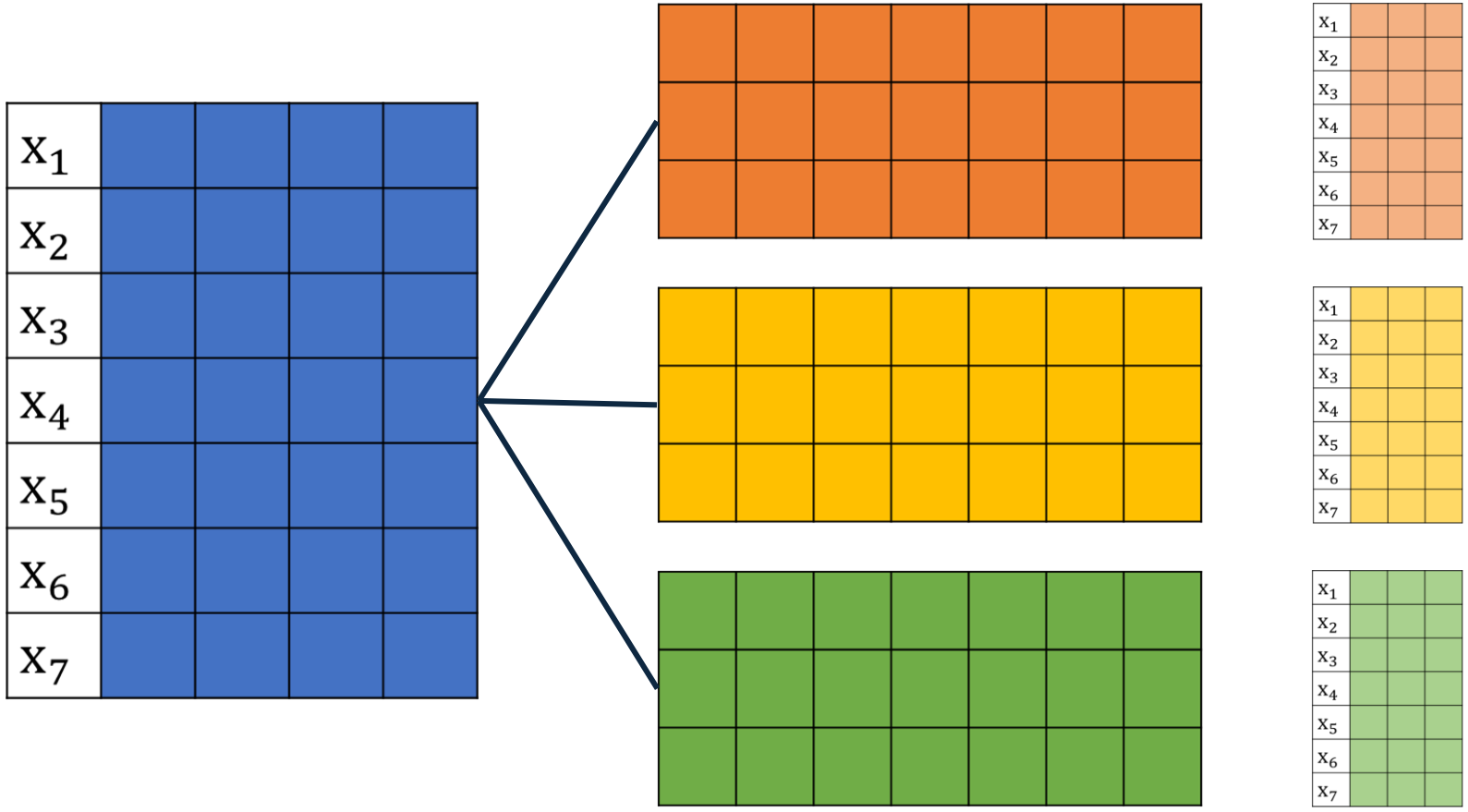
Cross attention: cross alignment

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.



Cross attention: cross alignment



$$\boldsymbol{X} \cdot \boldsymbol{W_Q} = \boldsymbol{Q}(\text{Query})$$

$$\boldsymbol{X} \cdot \boldsymbol{W_K} = \boldsymbol{K}(\text{Key})$$

$$\boldsymbol{X} \cdot \boldsymbol{W_V} = \boldsymbol{V}(\text{Value})$$

Daniluk, M., Rocktäschel, T., Welbl, J., & Riedel, S. (2017). Frustratingly short attention spans in neural language modeling. *arXiv preprint arXiv:1702.04521*.

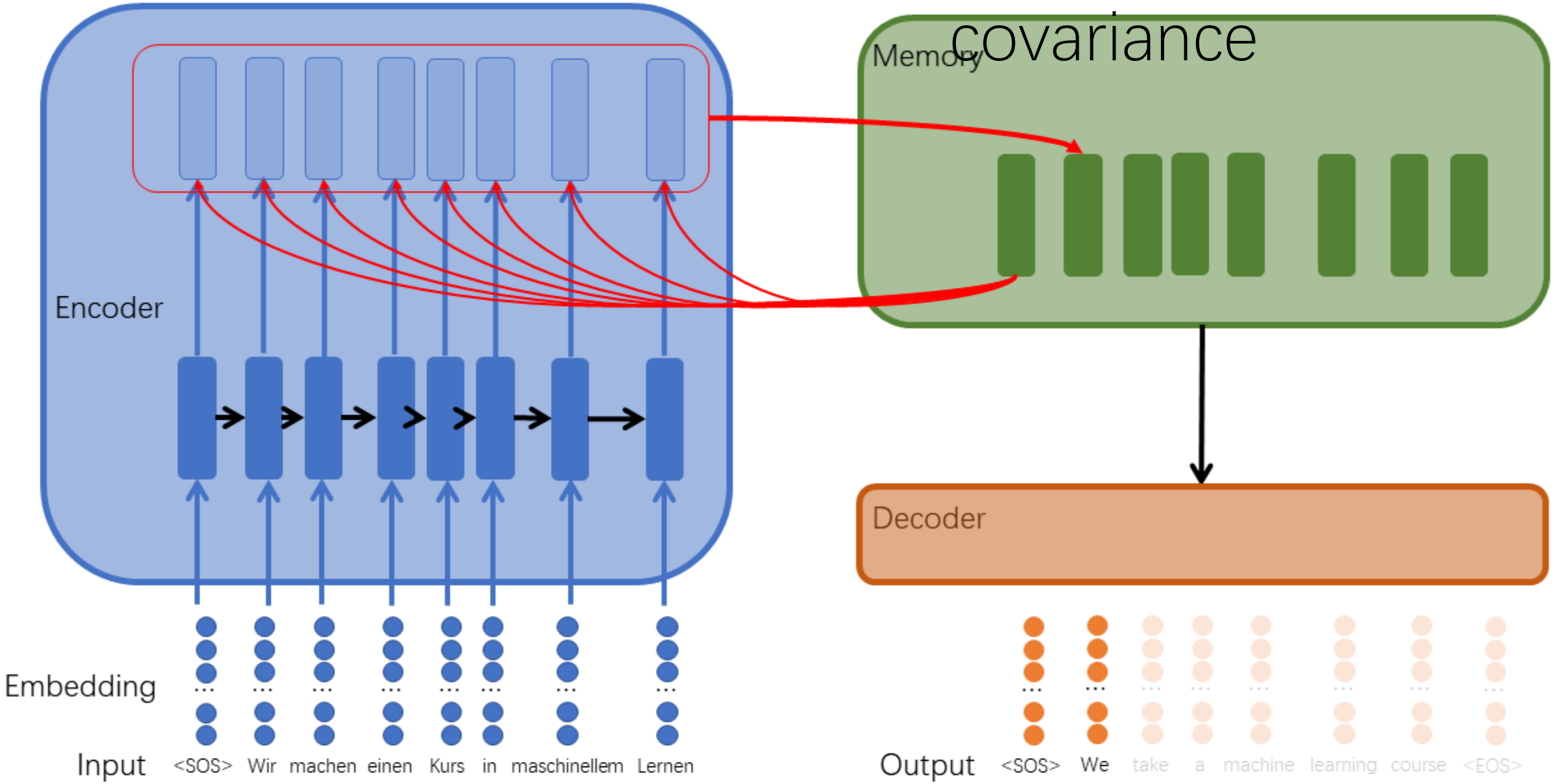
Cross attention: cross alignment

$$\text{Attention}(\boldsymbol{Q}_y, \boldsymbol{K}_x, \boldsymbol{V}_x) \rightarrow \boldsymbol{Z}$$

	We	take	a	machine	learning	course
Wir	0.8	0.18	0.01	0.01	0	0.01
machen	0.15	0.5	0.15	0.02	0.03	0.15
einen	0.01	0.15	0.8	0.03	0.004	0.006
Kurs	0.01	0.15	0.02	0.04	0.009	0.75
in	0.01	0.01	0.01	0	0.007	0.002
maschinellem	0.01	0.005	0.01	0.65	0.25	0.04
Lernen	0.01	0.005	0.01	0.25	0.7	0.04

Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Self-attention: covariance



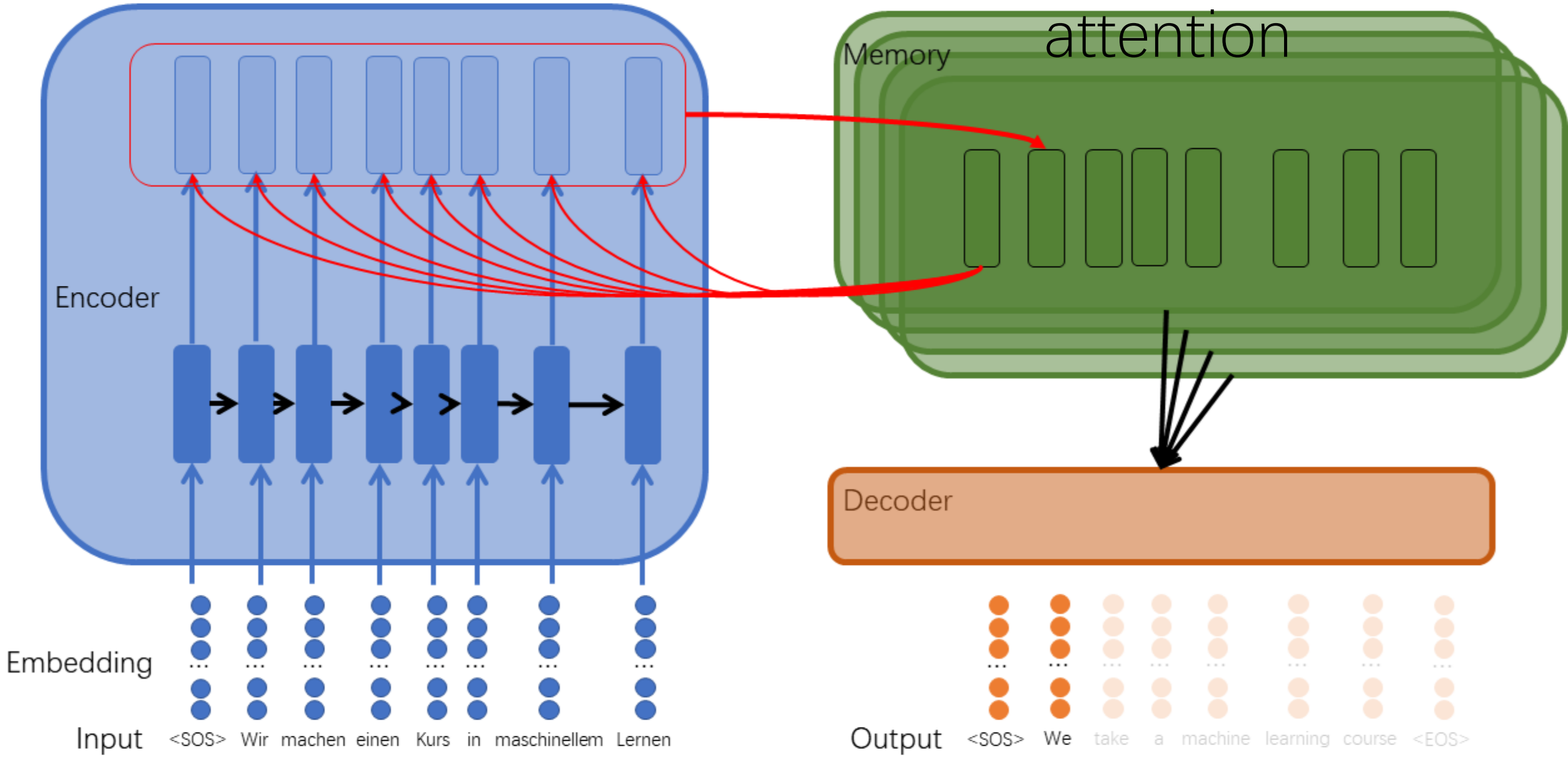
Self-attention: covariance

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

	We	take	a	machine	learning	course
We	0.6	0.35	0.02	0.01	0.01	0.01
take	0.35	0.5	0.4	0.01	0.02	0.2
a	0.02	0.4	0.6	0.01	0.02	0.04
machine	0.01	0.01	0.01	0.7	0.25	0.02
learning	0.01	0.02	0.02	0.25	0.8	0.05
course	0.01	0.2	0.4	0.02	0.05	0.65

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

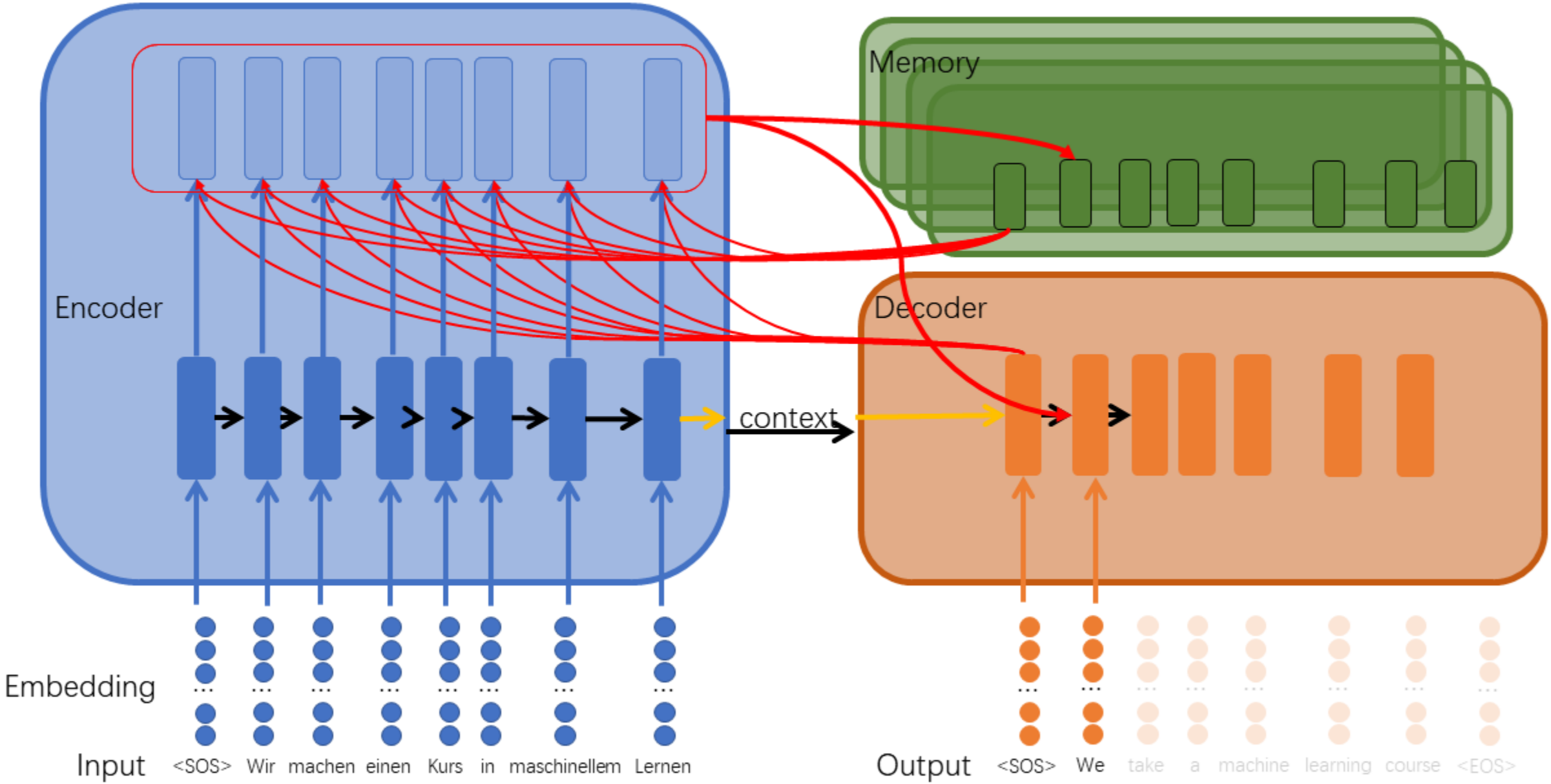
Multi-head attention



Multi-head attention: different informations

	We	take	a	machine	learning	course
We	0.5	0.2	0.02	0.01	0.01	0.35
take	0.25	0.5	0.02	0.01	0.02	0.2
a	0.02	0.02	0.3	0.01	0.02	0.04
machine	0.01	0.01	0.01	0.45	0.25	0.02
learning	0.01	0.02	0.02	0.35	0.5	0.22
course	0.4	0.2	0.4	0.02	0.3	0.65

	We	take	a	machine	learning	course
We	0.6	0.35	0.02	0.01	0.01	0.01
take	0.35	0.5	0.1	0.01	0.02	0.2
a	0.02	0.1	0.6	0.01	0.02	0.5
machine	0.01	0.01	0.01	0.7	0.25	0.02
learning	0.01	0.02	0.02	0.25	0.8	0.05
course	0.01	0.2	0.4	0.02	0.05	0.65

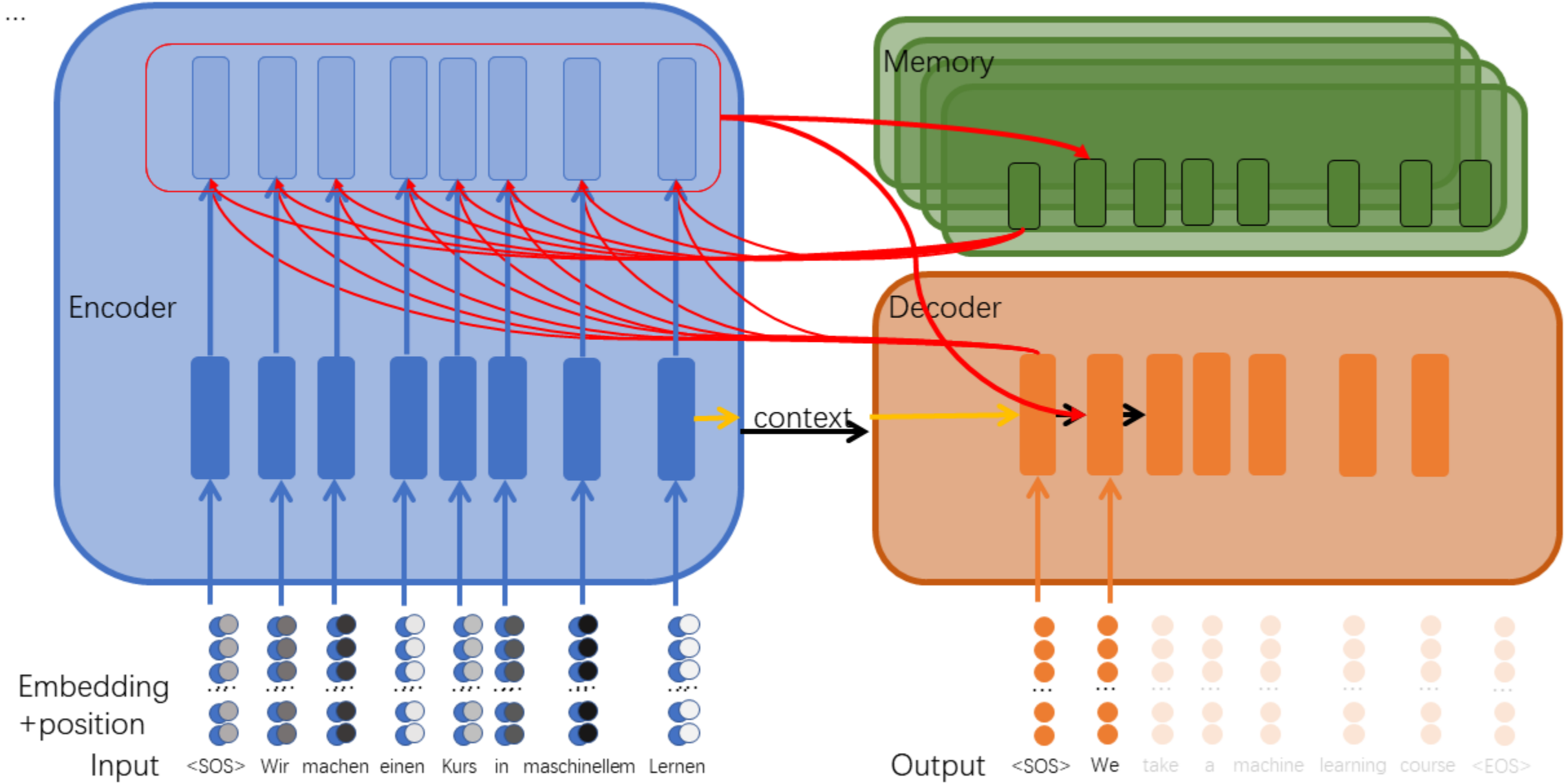

Memory
Encoder
Decoder
context
Embedding
Input
<SOS> Wir machen einen Kurs in maschinellem Lernen
Output
<SOS> We take a machine learning course <EOS>

Attention! please

- Cross attention - Self attention - Multi-head Attention

It’s time to get rid of RNN!

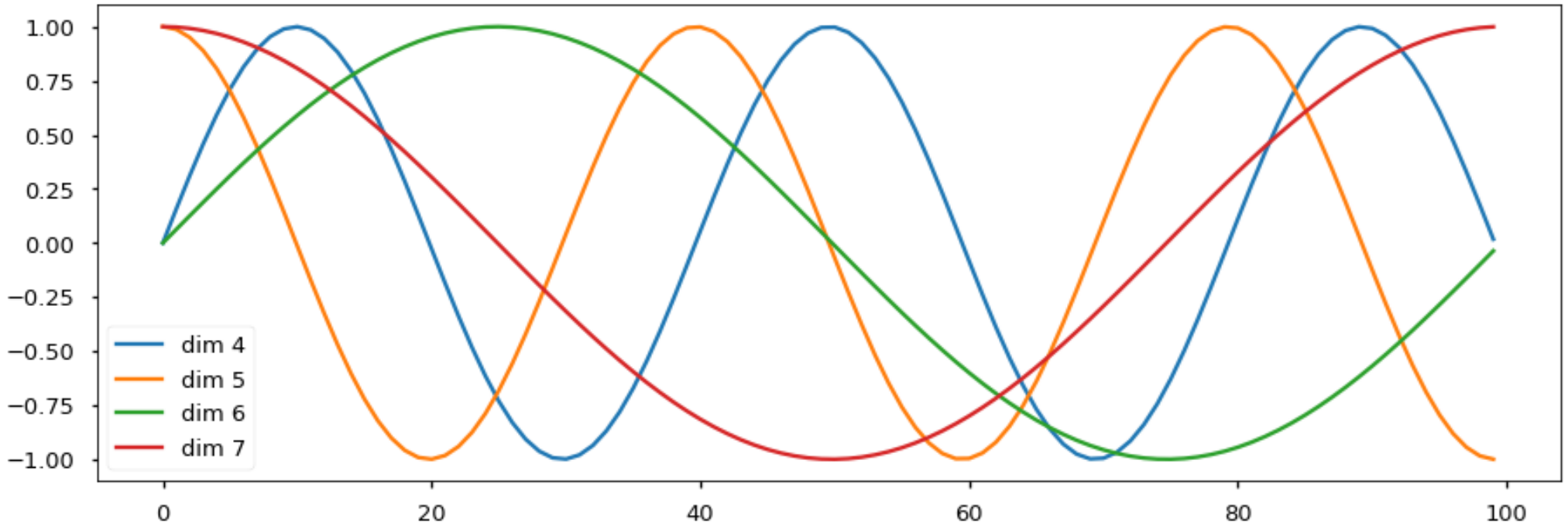
(sequential data…)


...
Memory
Encoder
Decoder
context
Embedding
+position
Input
<SOS> Wir machen einen Kurs in maschinellem Lernen
Output
<SOS> We take a machine learning course <EOS>

Position embedding

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d}\right)$$

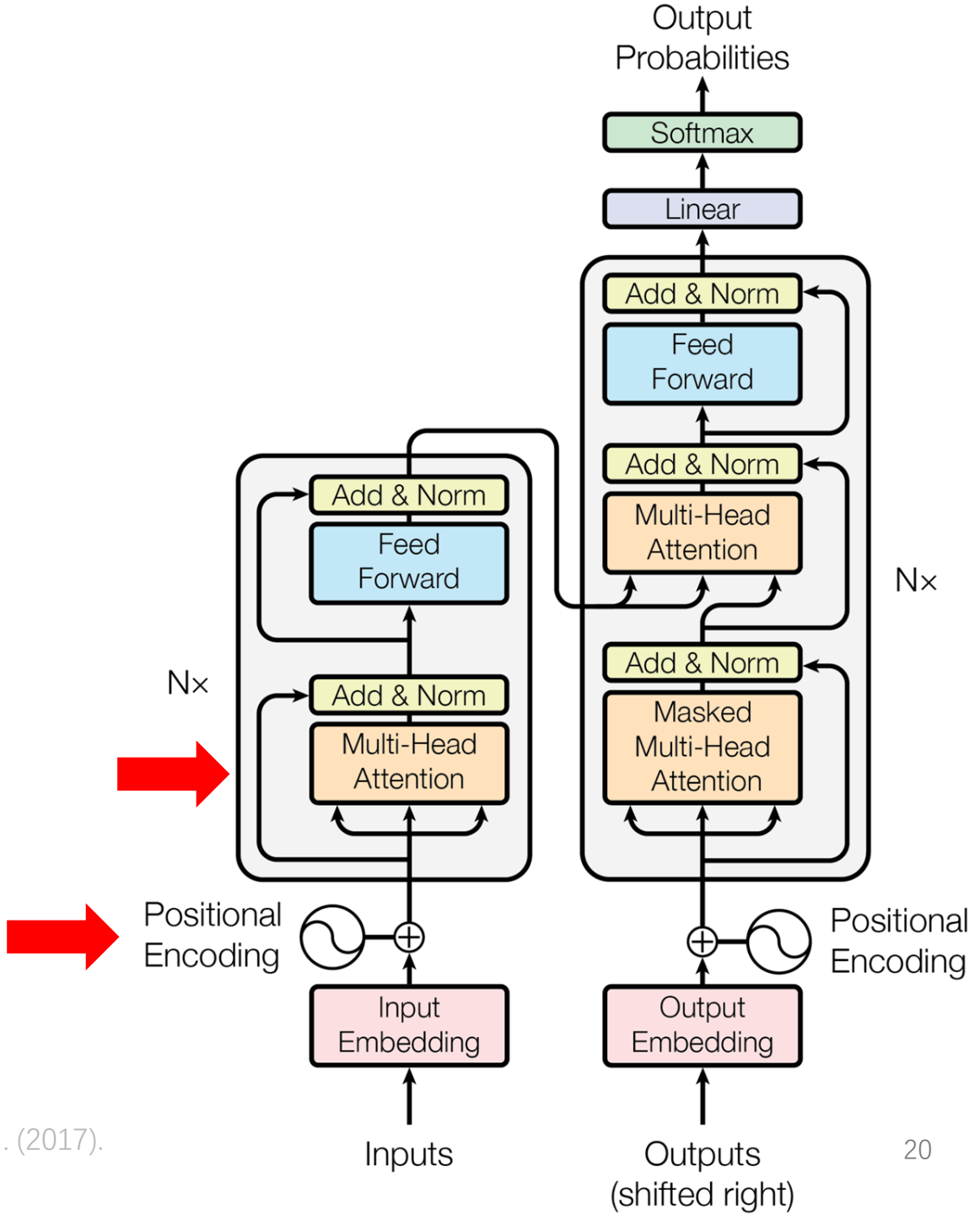


Attention is all you need

Encoder:

Embedding and position embedding

Self-attention / multi-head



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

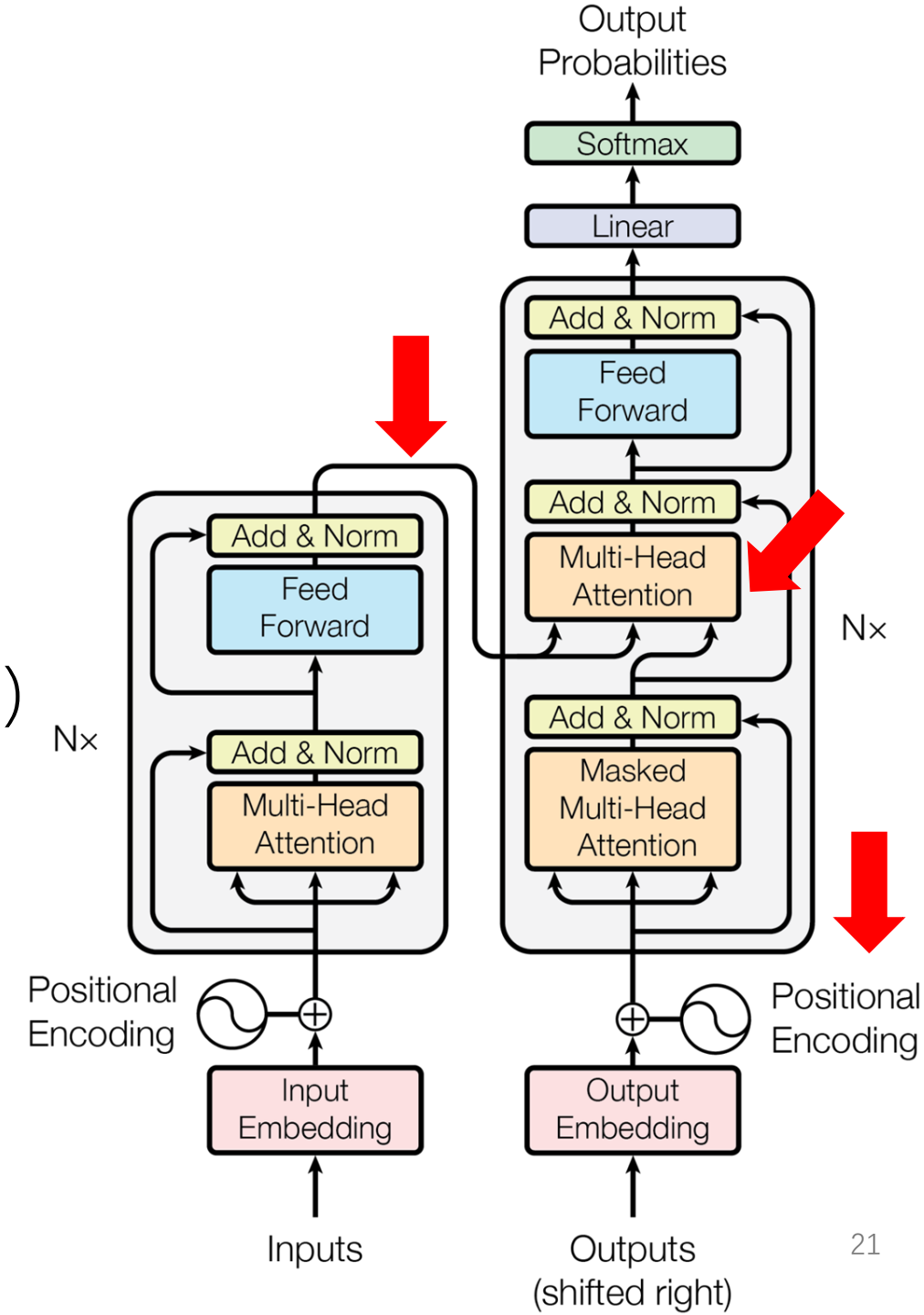
Attention is all you need

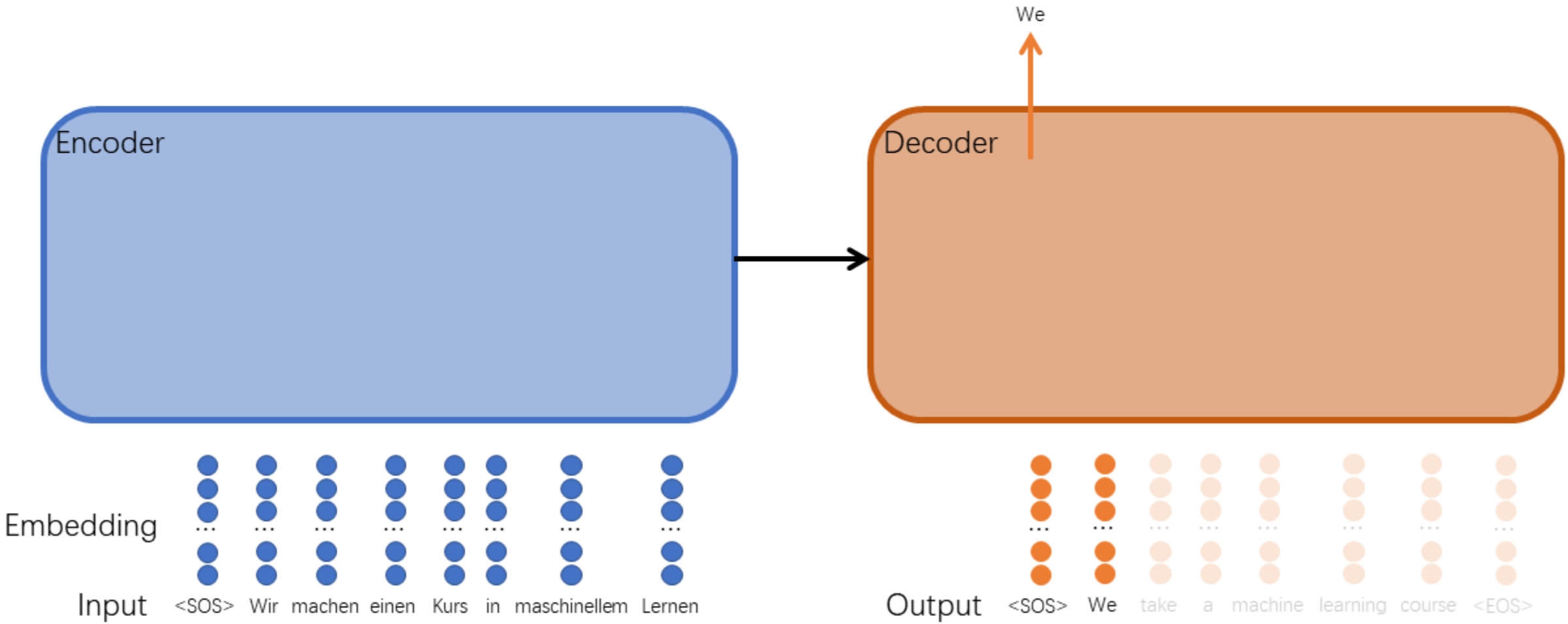
Decoder:

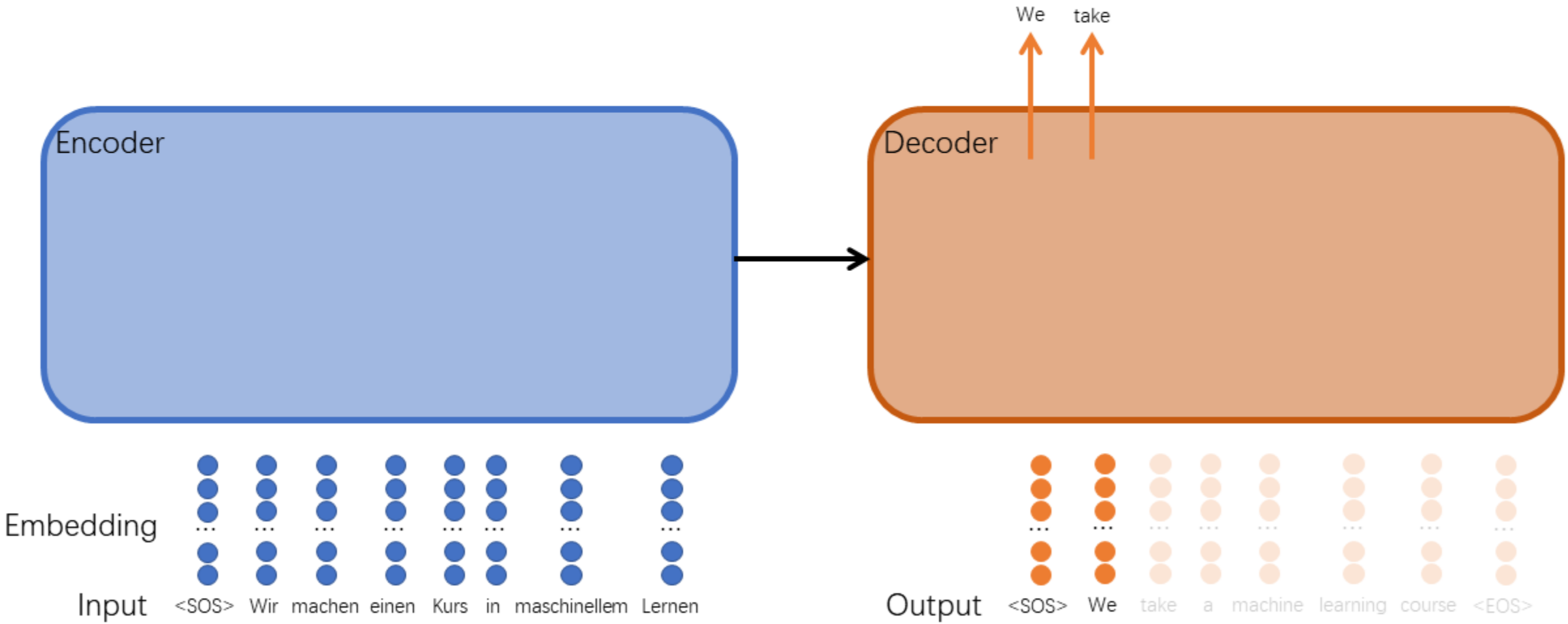
Embedding and position embedding

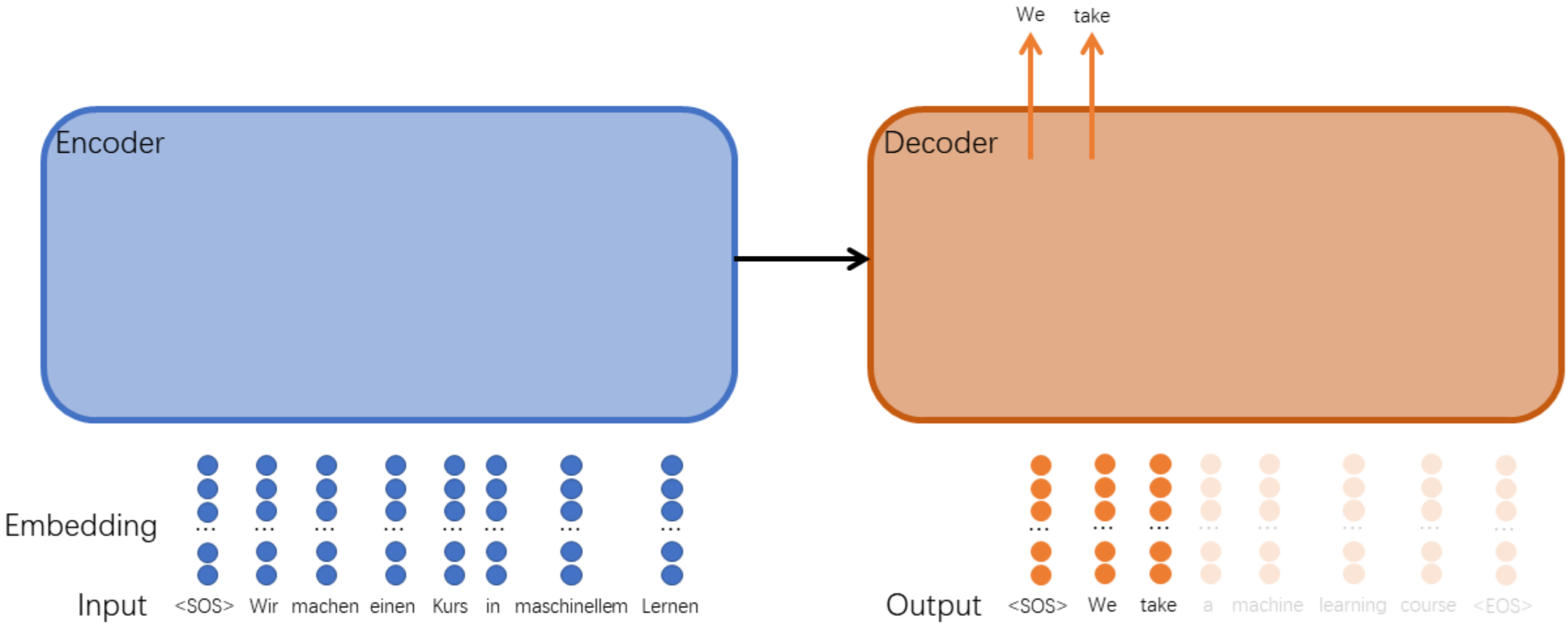
Self-attention / multi-head/ mask(in training)

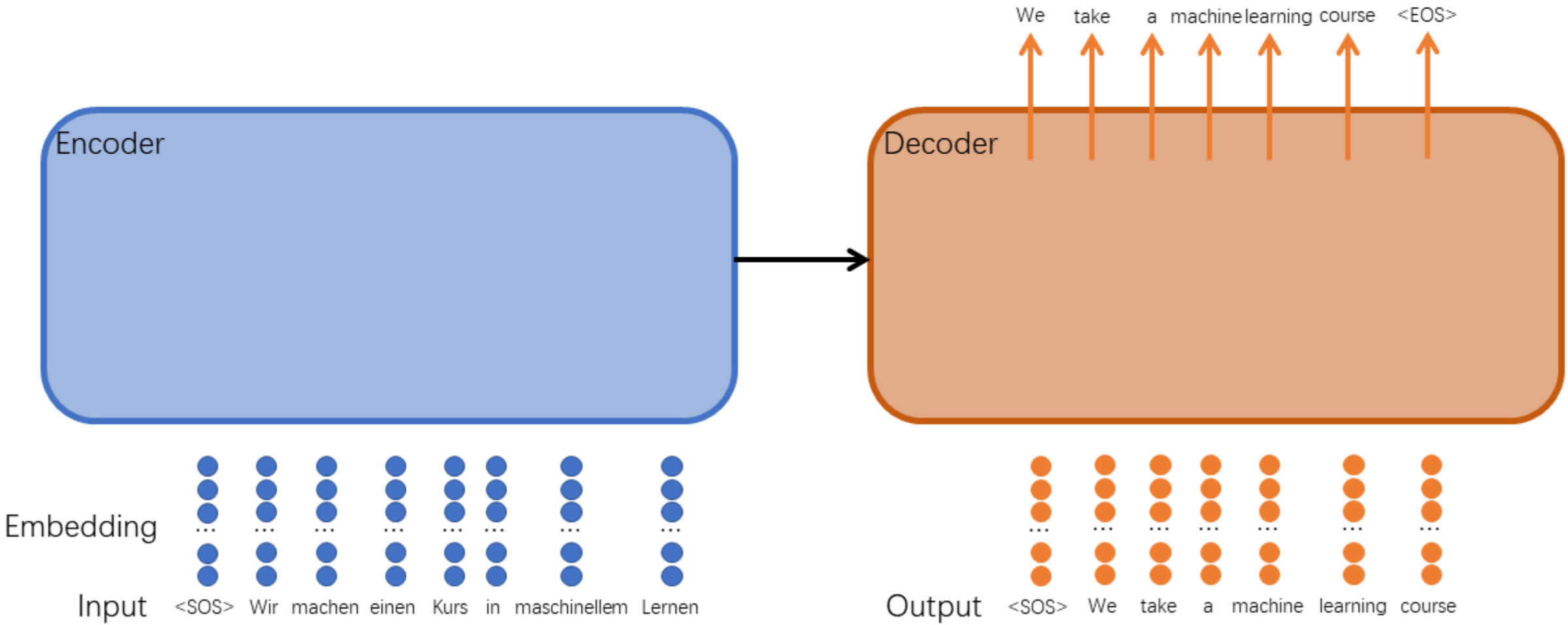
Cross-attention




We
Encoder
Decoder
Embedding
Input
<SOS>
Wir
machen
einen
Kurs
in
maschinellem
Lernen
Output
<SOS>
We
take
a
machine
learning
course
<EOS>


We
take
Encoder
Decoder
Embedding
Input
<SOS>
Wir
machen
einen
Kurs
in
maschinellem
Lernen
Output
<SOS>
We
take
a
machine
learning
course
<EOS>


We
take
Encoder
Decoder
Embedding
Input
<SOS>
Wir
machen
einen
Kurs
in
maschinellem
Lernen
Output
<SOS>
We
take
a
machine
learning
course
<EOS>


We take a machine learning course <EOS>
Encoder
Decoder
Embedding
Input
<SOS> Wir machen einen Kurs in maschinellem Lernen
Output
<SOS> We take a machine learning course

The Annotated Transformer

http://nlp.seas.harvard.edu/annotated-transformer/

GitHub:harvardnlp/annotated-transformer: An annotated implementation of the Transformer paper.