NeurIPS 2018

Layer-Wise Coordination between Encoder and Decoder for Neural Machine Translation

Tianyu He[1 † *]
hetianyu@mail.ustc.edu.cn

Xu Tan[2 †]
xuta@microsoft.com

Yingce Xia[2]
yingce.xia@microsoft.com

Di He[3]
di_he@pku.edu.cn

Tao Qin[2]
taoqin@microsoft.com

Zhibo Chen[1]
chenzhibo@ustc.edu.cn

Tie-Yan Liu[2]
tie-yan.liu@microsoft.com

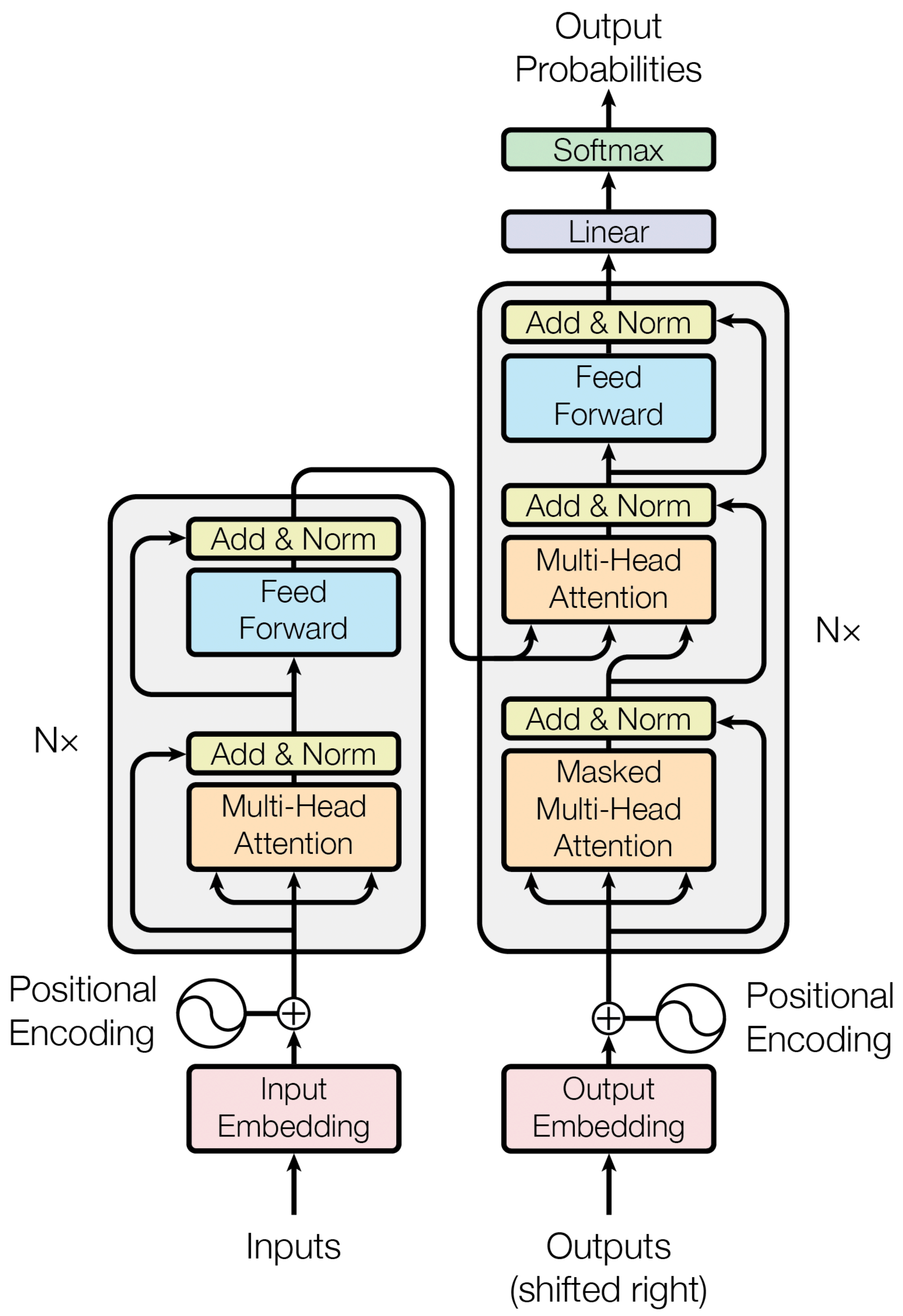
[1]CAS Key Laboratory of Technology in
Geo-spatial Information Processing and Application System,
University of Science and Technology of China
[2]Microsoft Research
[3]Key Laboratory of Machine Perception, MOE, School of EECS, Peking University

Reporter: Shen Sijie

Motivation

- Hidden states of target tokens are all generated from **highest-level** representations of the source sentence.

- Why should the low-level representation of a target token base on the highest-level ones of source tokens?



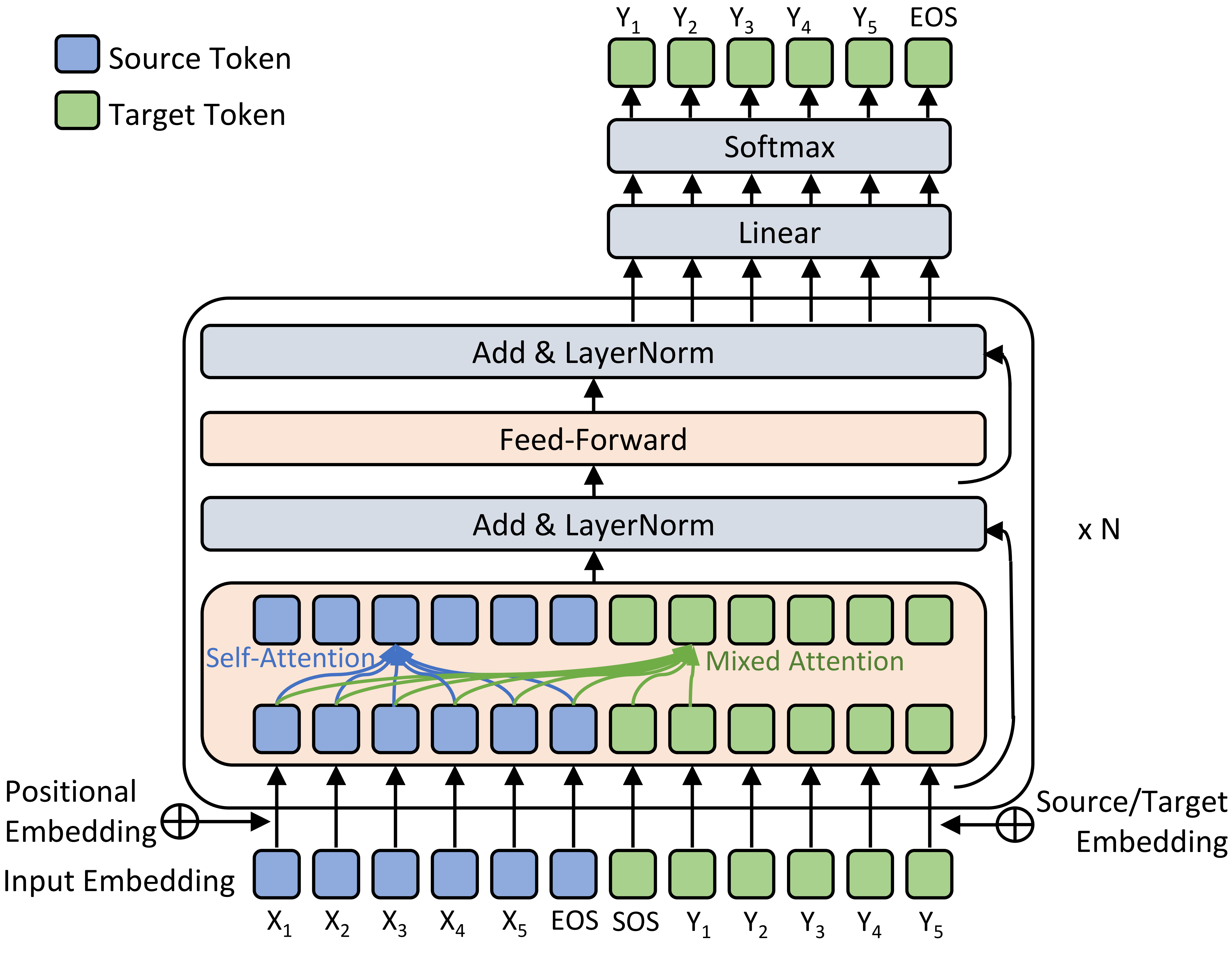
Layer-Wise Coordination

- Mechanism:
  - Hidden states in the i-th layer of decoder is generated from $(i-1)$-th layer of encoder and decoder.
  - Share the parameters of the encoder and decoder.
- This idea can be applied to many architectures:
  - RNN, CNN, Transformer

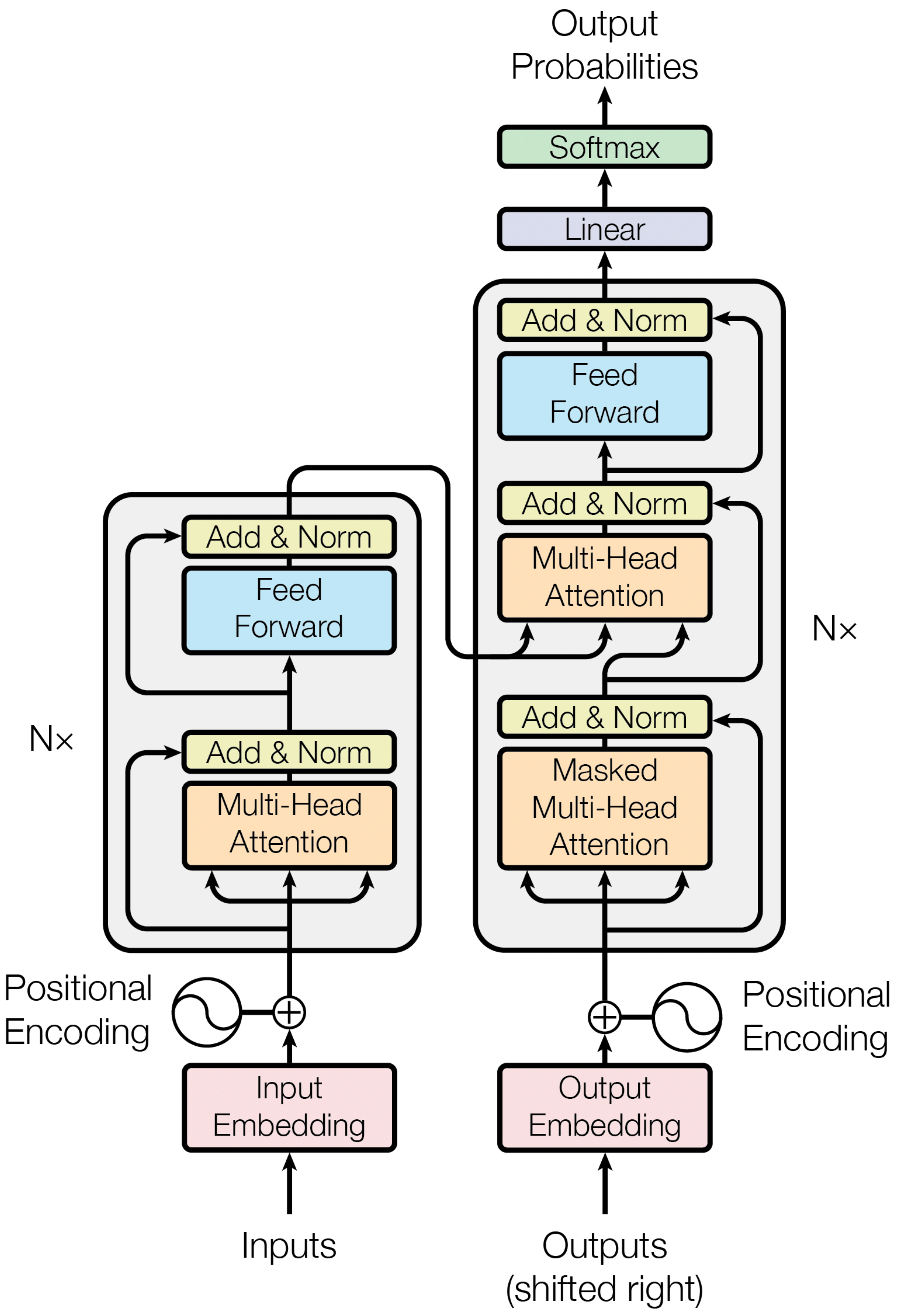
Layer-Wise Coordination

- Advantages:
  - The information from the source and target sentence will meet earlier, starting from the low-level representations.
  - Corresponding layers of the encoder and decoder are in the same (or closely related) semantic level.

Model



Layer-wise coordination



Original version of Transformer

Model

- Modifications:
  - Layers of decoder attends to corresponding layers of encoder.
  - Use mixed attention instead of encoder-decoder attention.
  - Share parameters of attention and feed-forward layer between encoder and decoder.
  - Use the sum of input embedding, source/target embedding and positional embedding as word representation.

Model

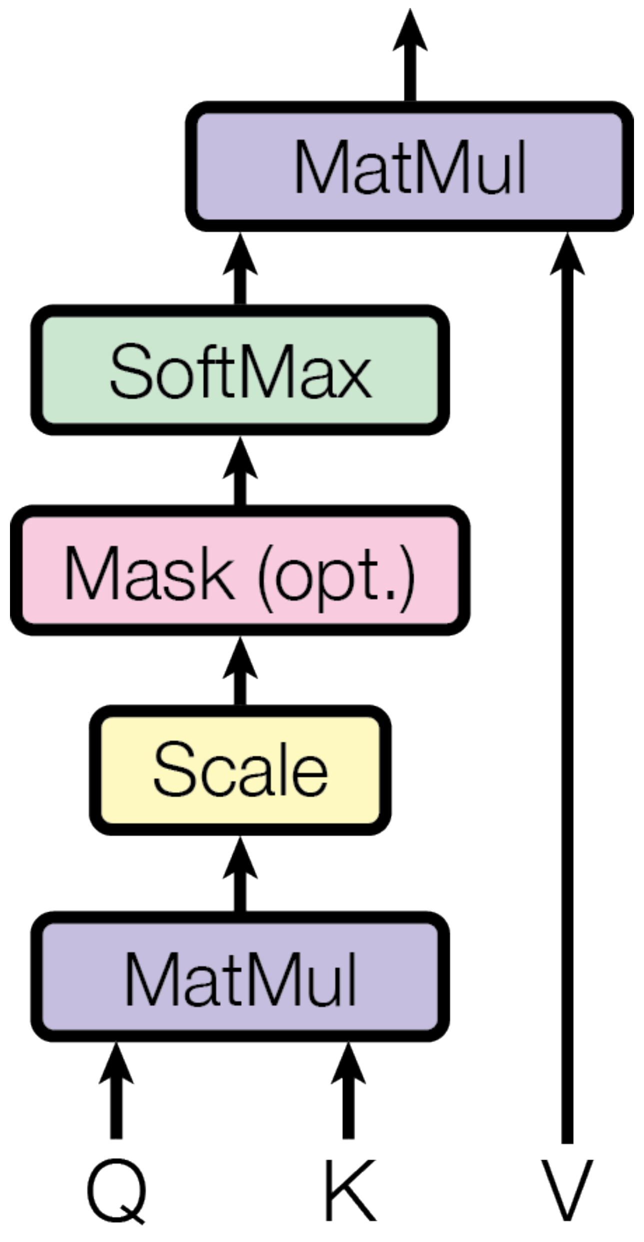
- **Mixed Attention:**

$$\text{Mixed\_Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_{\text{model}}}} + M)V,$$

$$M(i,j) = \begin{cases} 0, & j < n \vee j \leq i+n \\ -\infty, & \text{otherwise} \end{cases},$$

- Compared to original version of attention:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_{\text{model}}}})V,$$



Model

- **Position Embedding:** The same as Transformer.

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

- **Source/Target Embedding:**
  - Use two embeddings for source and target language respectively.
  - Learned in training process.

Experiment

- Dataset:
  - IWSLT14 German/Romanian/Spanish-English (De-En/Ro-En/Es-En)
  - WMT16 English-Romanian (En-Ro)
  - WMT14 English-German (En-De)
- Configuration:

Small: $d_{model} = 256, d_{ff} = 1024$

Base: $d_{model} = 512, d_{ff} = 2048$

Big: $d_{model} = 1024, d_{ff} = 2048$

Experiment

- Result:

Task	Method	BLEU
De-En	MIXER [23]	21.83
	AC+LL [1]	28.53
	NPMT [11]	28.96
	Dual Transfer Learning [34]	32.35
	Transformer (small)	32.86
	Our method (small)	**35.07**
Ro-En	Transformer(small)	29.64
	Our method (small)	**30.72**
Es-En	UEDIN[3]	37.29
	Transformer(small)	38.57
	Our method (small)	**40.50**

Table 1: BLEU scores on IWSLT 2014 translation tasks compared with transformer baseline and other RNN/CNN-based models.

Task	Method	BLEU
En-Ro	GRU[24]	28.10
	ConvS2S[7]	30.02
	Transformer (big)	32.70
	Our method (big)	**34.43**
En-De	ByteNet [12]	23.75
	GNMT+RL [36]	24.60
	ConvS2S [7]	25.16
	MoE [26]	26.03
	Transformer (base) [33]	27.30
	Transformer (big) [33]	28.40
	Our method (base)	28.33
	Our method (big)	**29.01**

Table 2: BLEU scores on WMT translation tasks compared with transformer baseline and other RNN/CNN-based models.

Model variations

- Ablation Study:

	#parameter	BLEU	Δ
Our model	19.07M	35.07	
Our model w/o weight sharing	19.07M	33.96	1.11 $\downarrow$
Our model w/o mixed attention	19.07M	33.77	1.30 $\downarrow$
Our model w/o source/target embedding	19.07M	32.80	2.27 $\downarrow$
Our model w/o position embedding	19.07M	18.46	16.61 $\downarrow$

Table 3: Ablation study on our proposed model on De-En task.

- Varying the Number of Layers:

#layer	10	14	18	22	#layer	4	6	8	10
Our method	34.32	35.07	35.31	35.05	Baseline	32.78	32.86	32.72	32.67

Table 4: The BLEU scores under different number of layers for our method and the baseline on De-En task.

Case Study

- Case Analysis:

Source (De)	zwei minuten später passierten drei dinge gleichzeitig.
Reference (En)	two minutes later, three things happened at the same time.
Transformer	two minutes later, three things happened.
Our model	two minutes later, three things happened at the same time.
Source (De)	mit 17 wurde sie die zweite frau eines mandarin, dessen mutter sie schlug.
Reference (En)	at 17 she became the second wife of a mandarin whose mother beat her.
Transformer	at the age of 17, she turned into a mandarin second woman whose mother beat her.
Our model	at 17, she became the second woman of a mandarin whose mother beat her.
Source (De)	und ich erwiderte: "wie kommuniziert ihr denn nun?"
Reference (En)	and i said, "well, how do you actually communicate?"
Transformer	and i said, "how does you communicates?"
Our model	and i said, "how do you communicate?"

Case Study

- Attention Visualization:

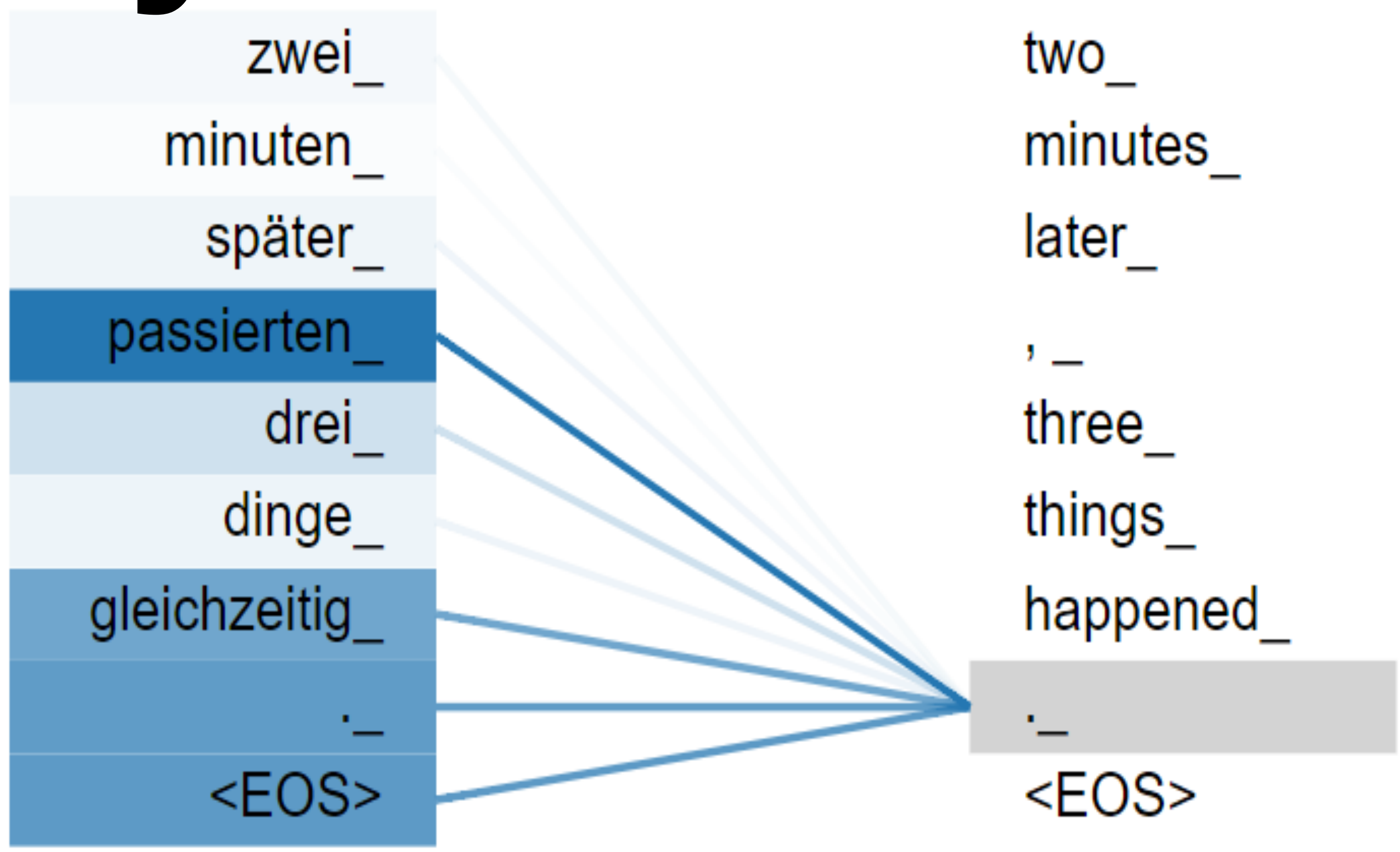


Figure 2: Source to target attention in Transformer.

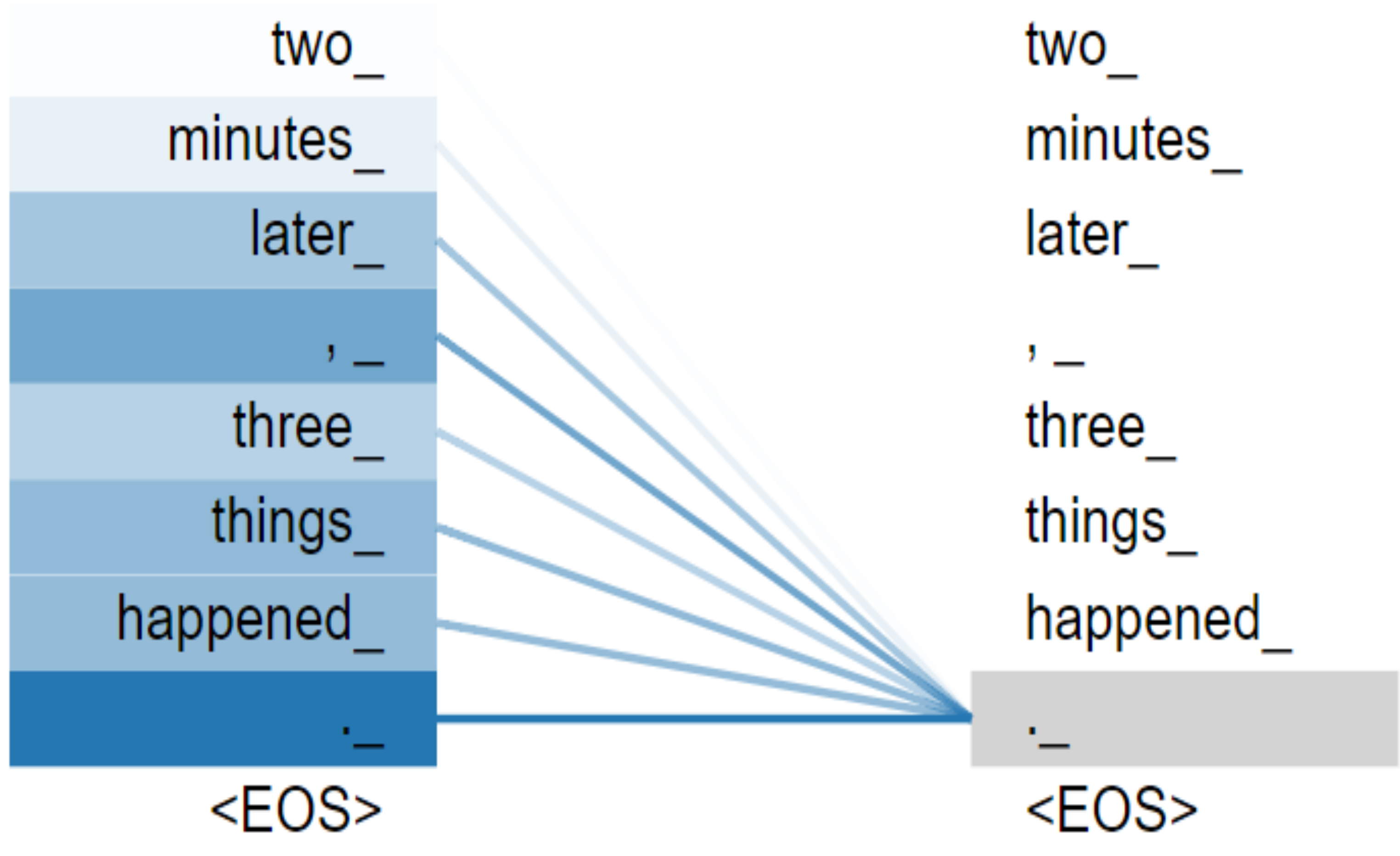


Figure 3: Target self-attention in Transformer.

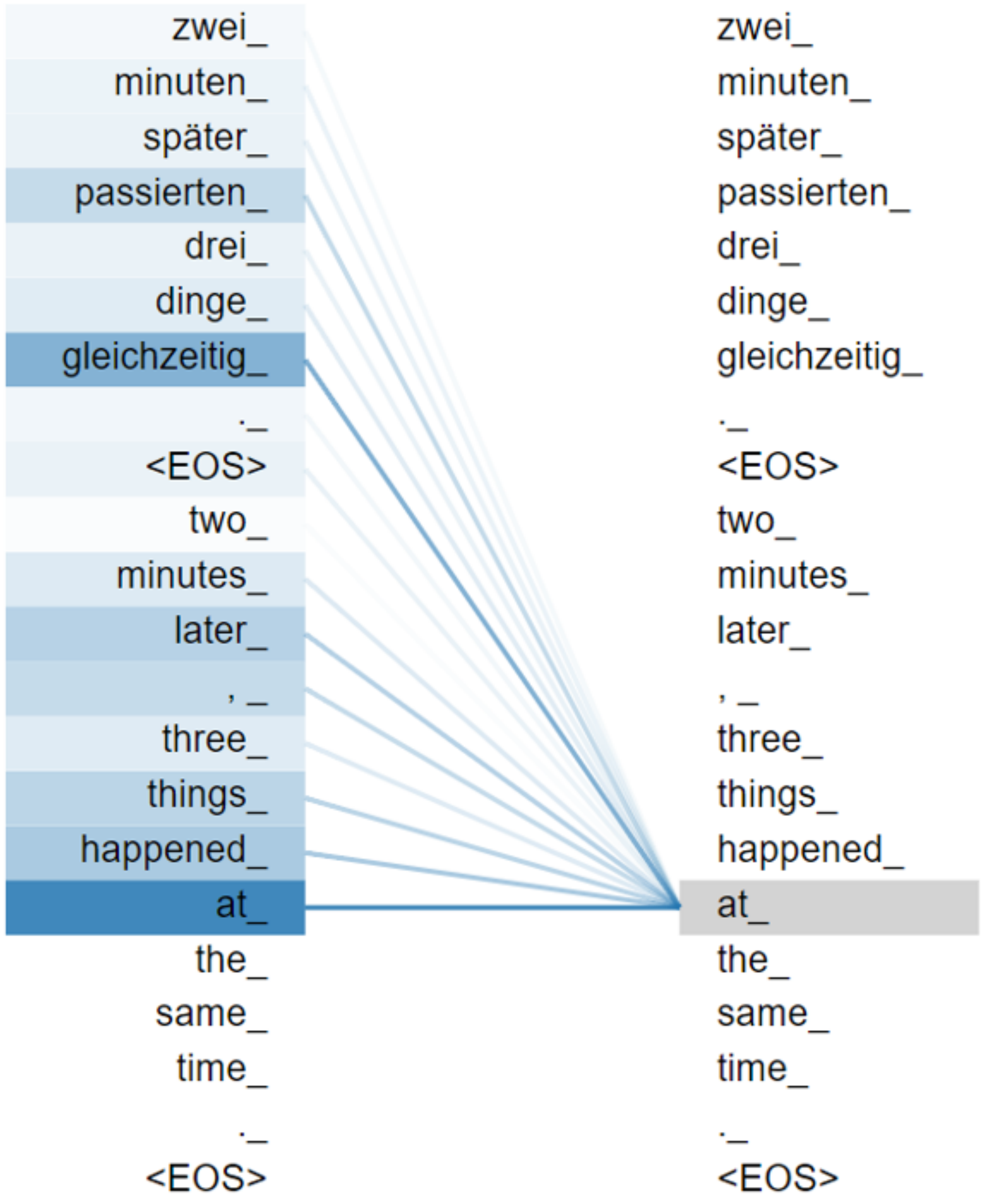


Figure 4: Mixed attention in our model.

Discussions on Mixed Attention

- In NMT:
  - Source contexts affect adequacy
  - Target contexts affect fluency
- Mixed Attention:
  - Automatically learns the preference on the source or target contexts

Thanks!