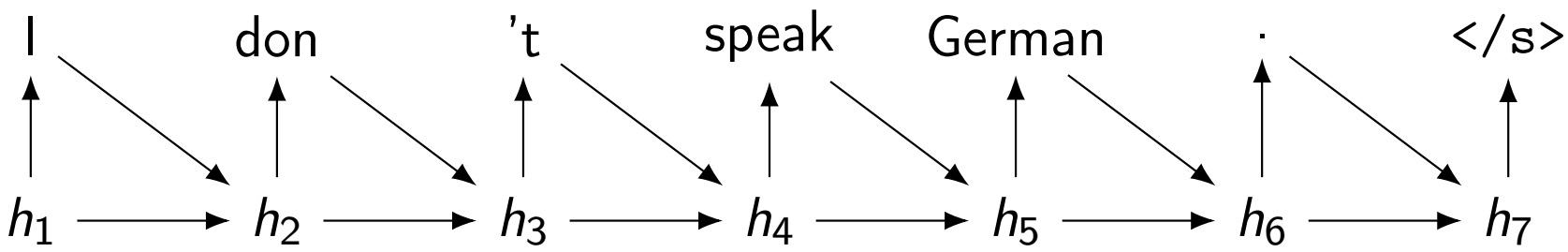
Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit,
Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

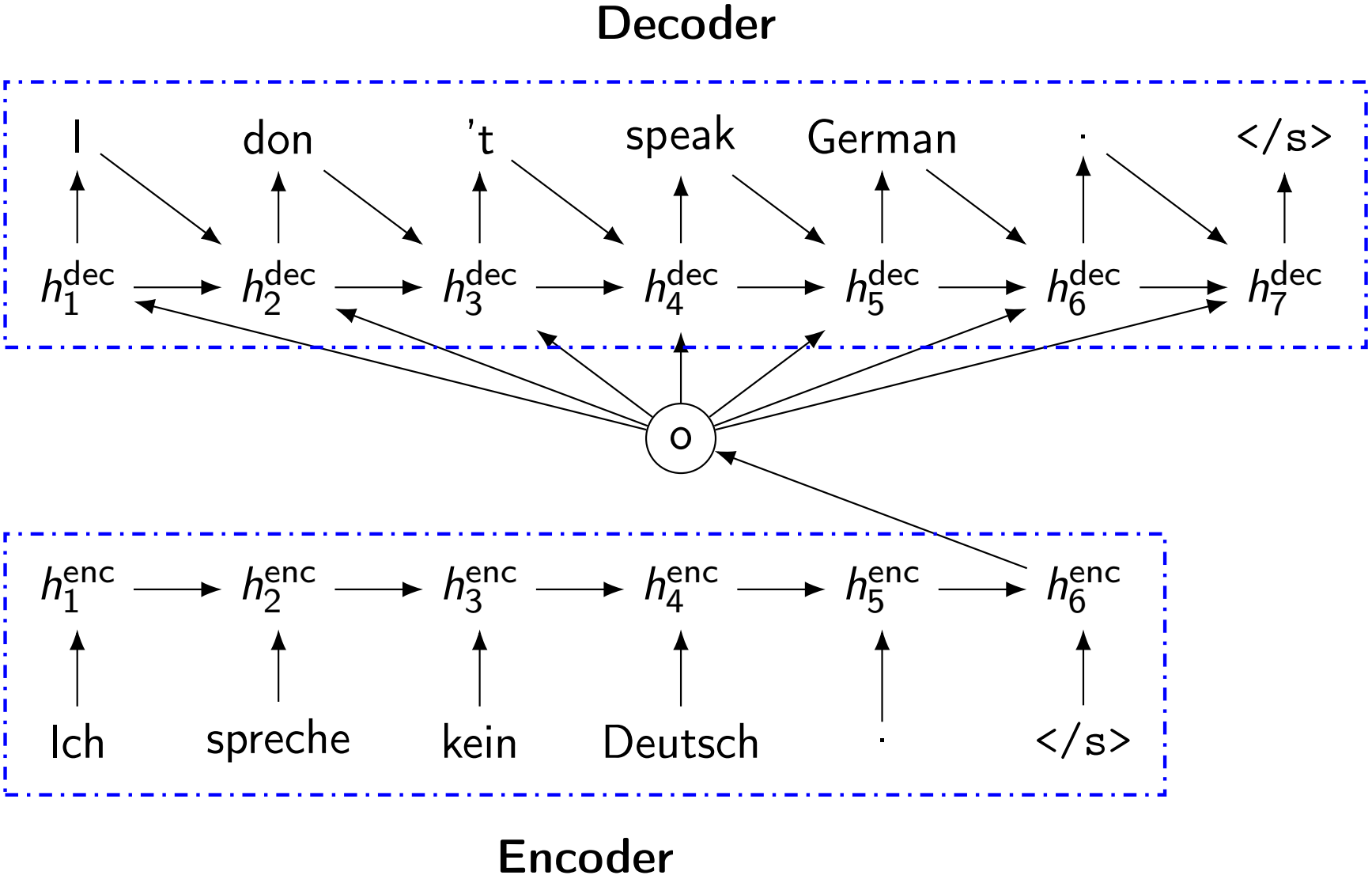
Google {Brain, Research}

June 2017

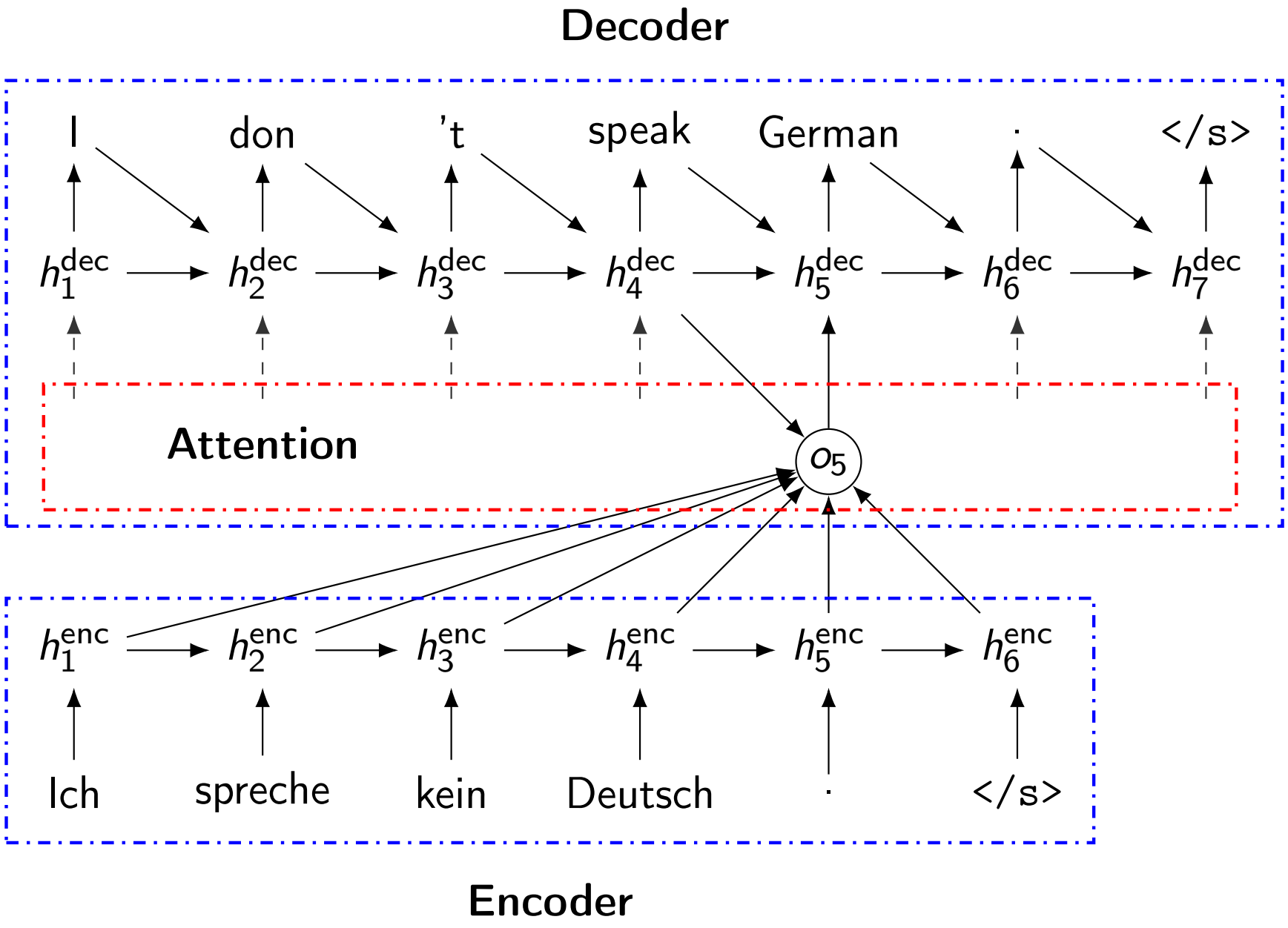
Neural Language Model



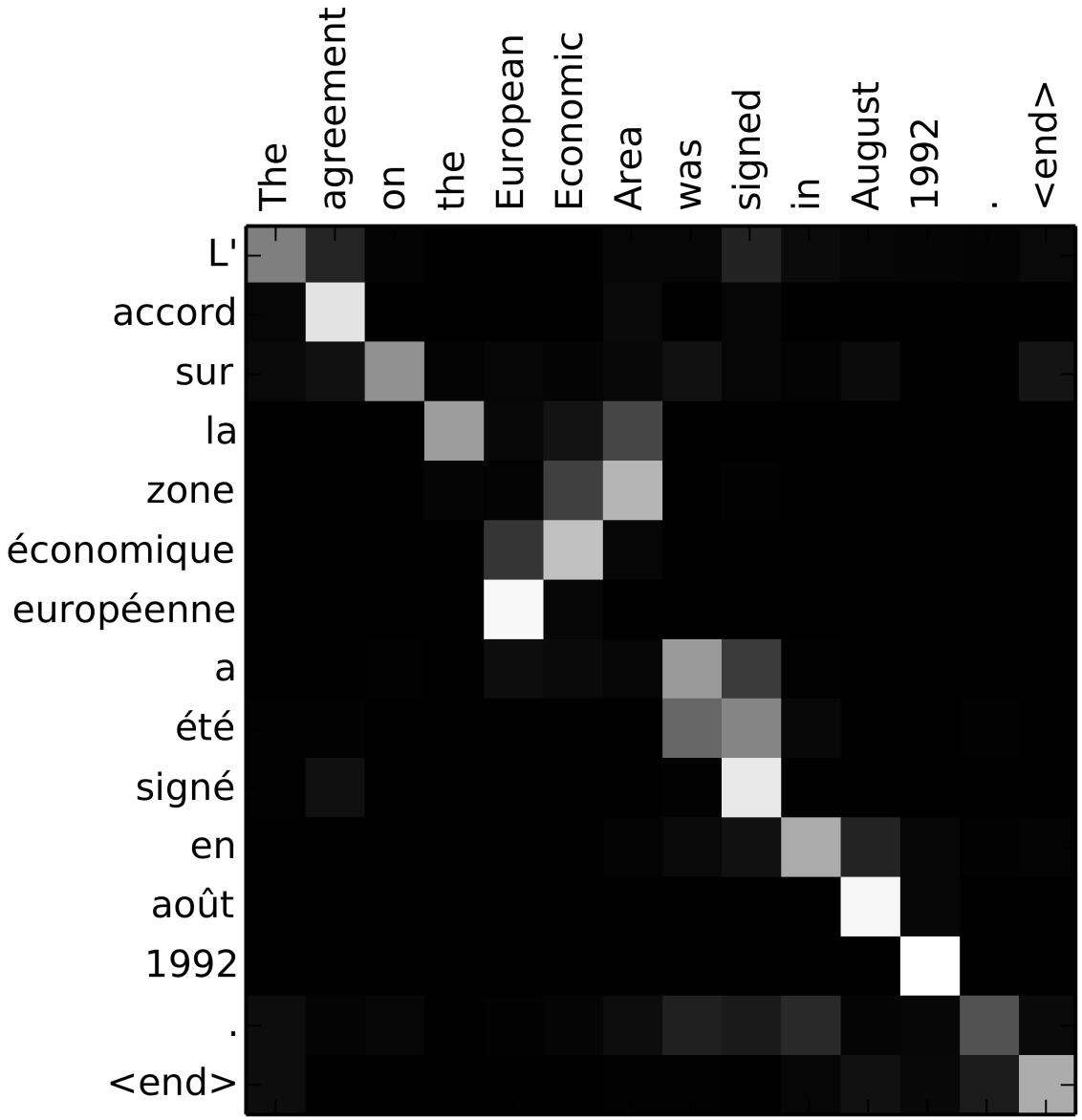
NMT Encoder-Decoder Model (Cho 2014)



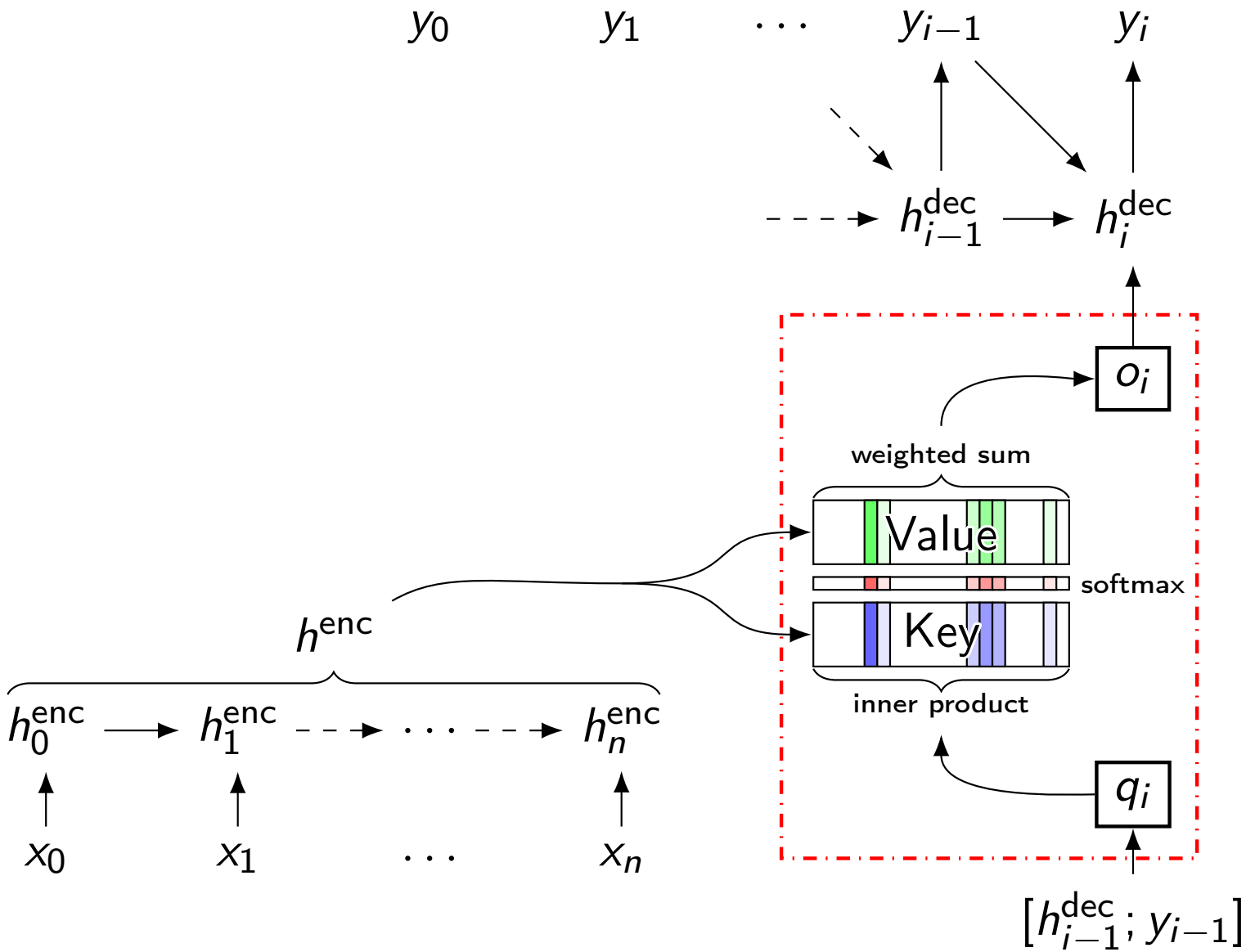
Decoder with Attention (2014 Bahdanau)



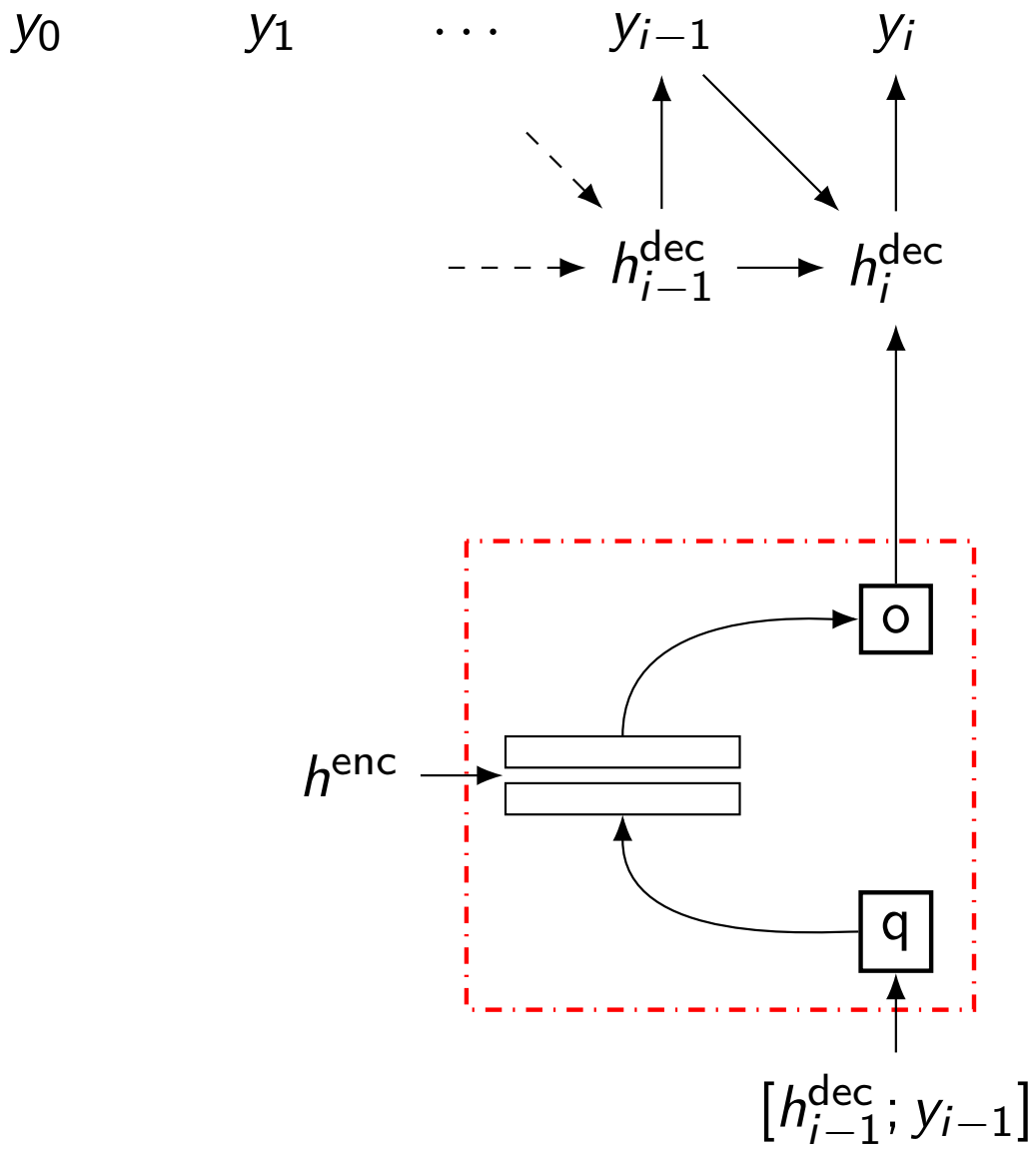
Attention Matrix



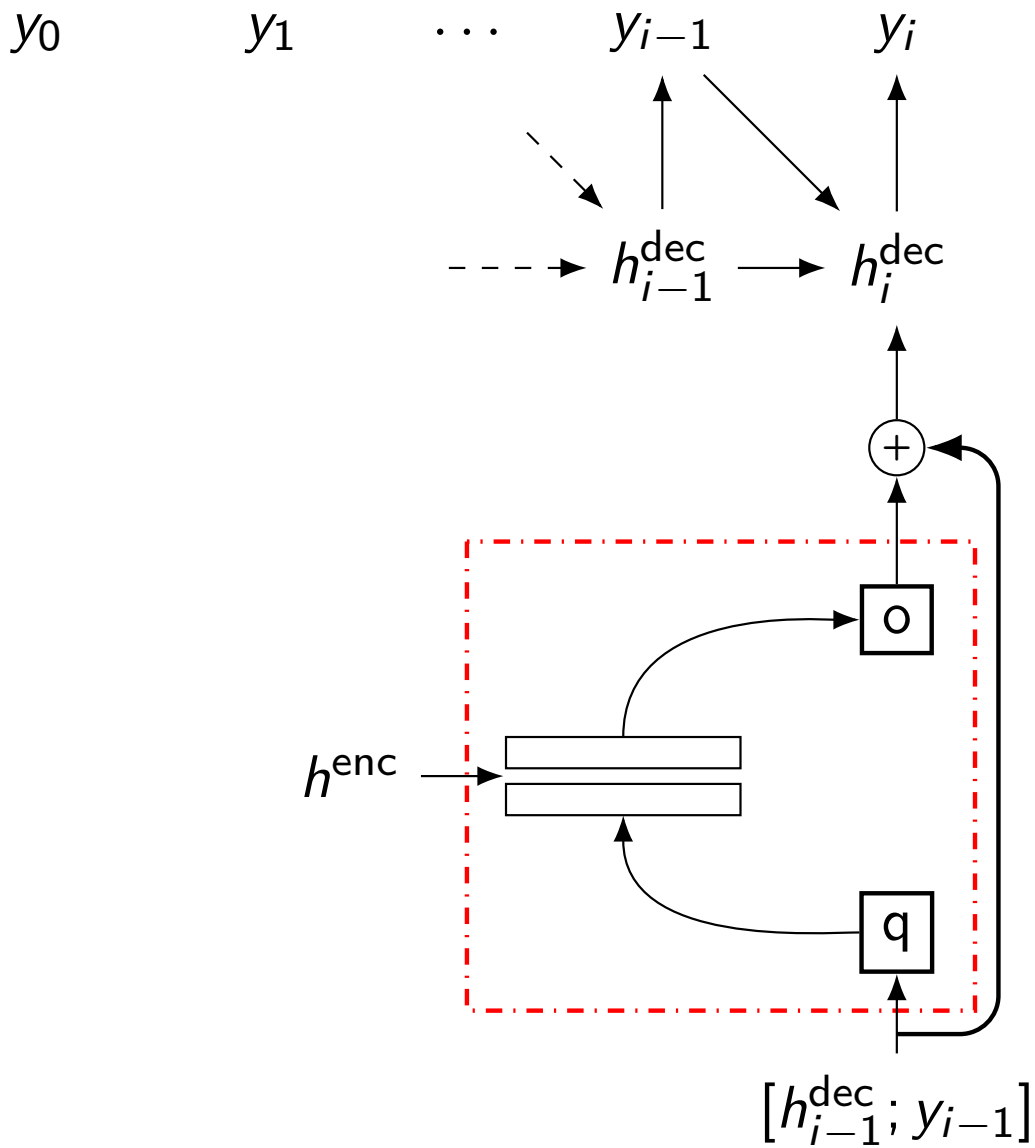
Attention as a Memory Network



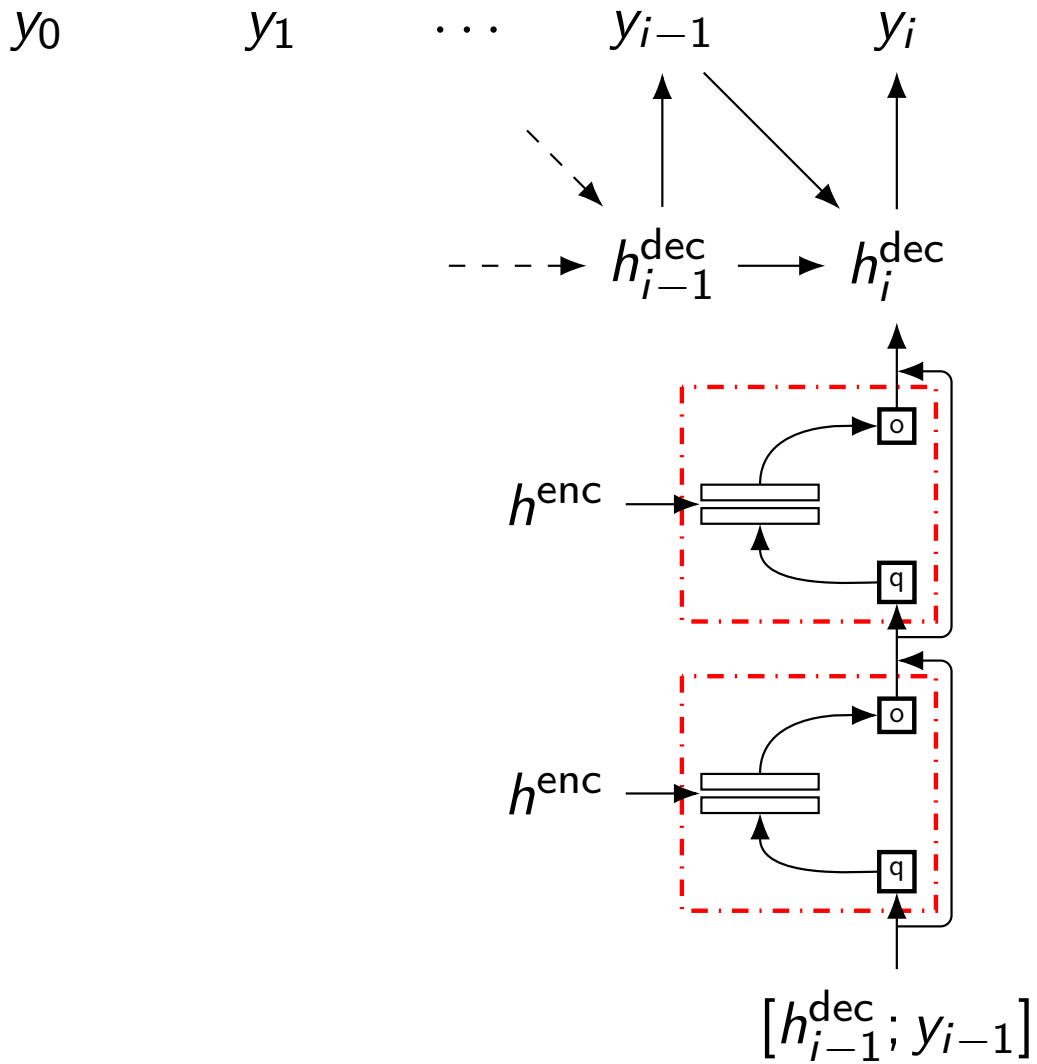
Decoder with Attention



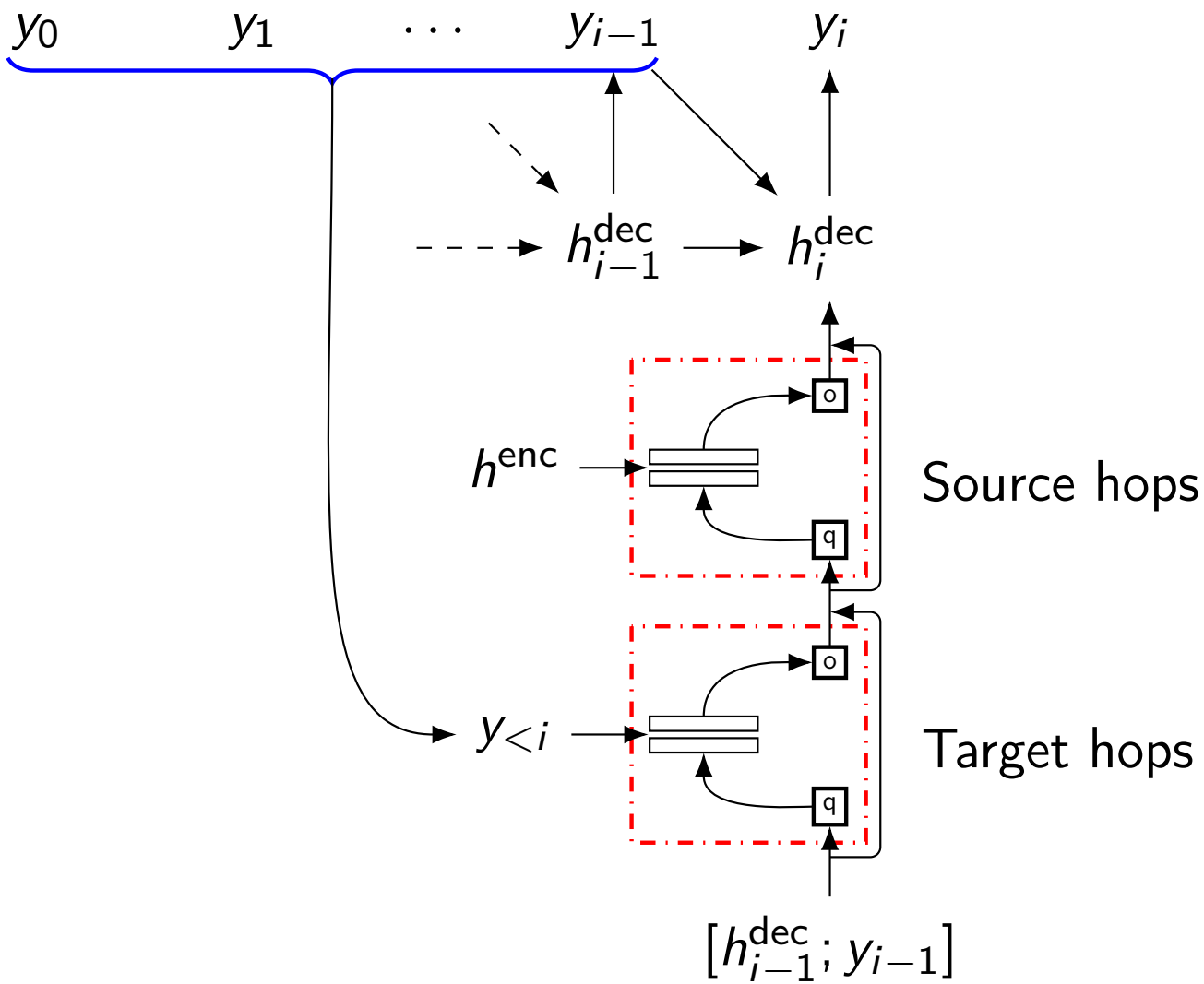
Residual Connection



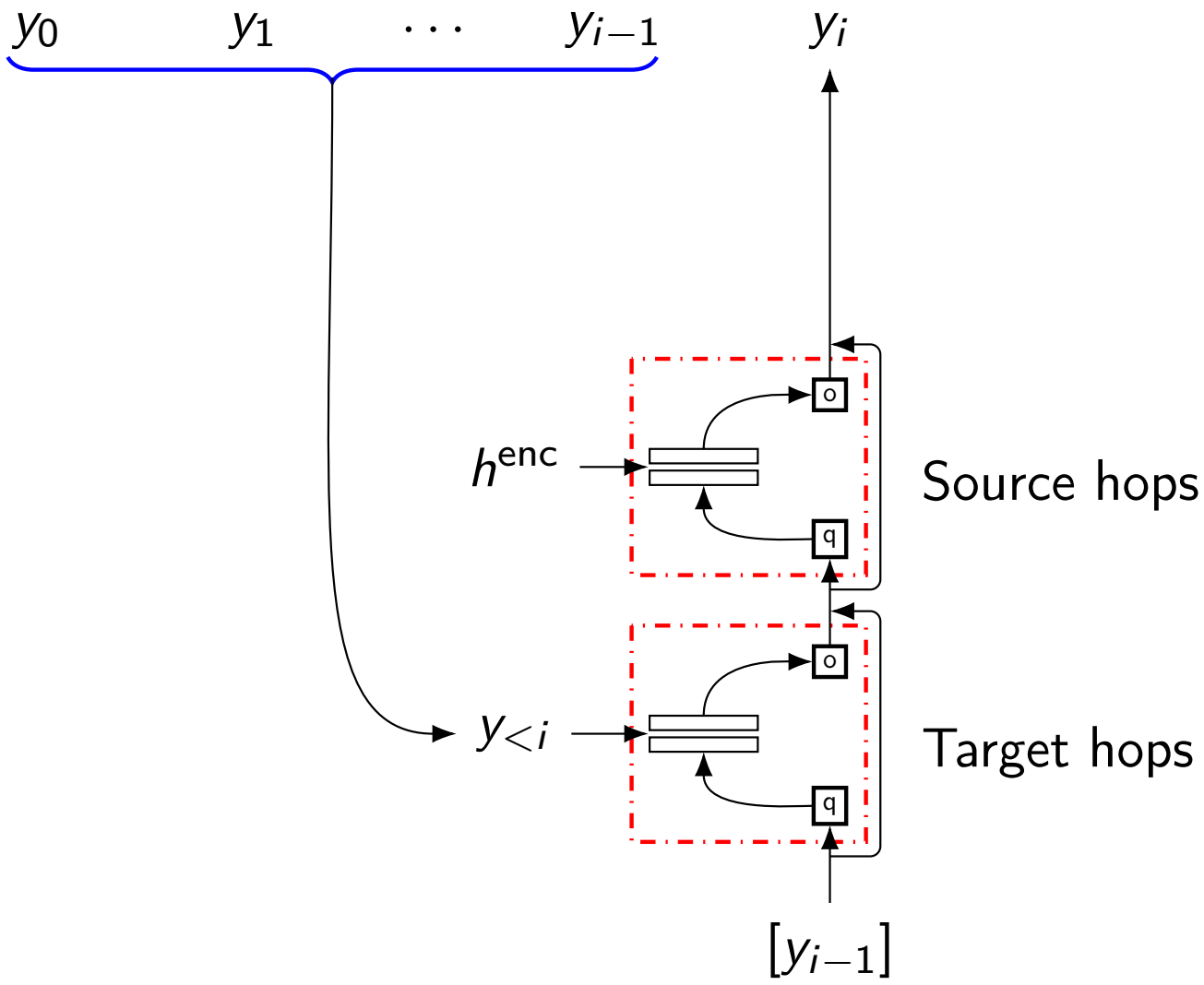
Decoding with Multiple Source Hops



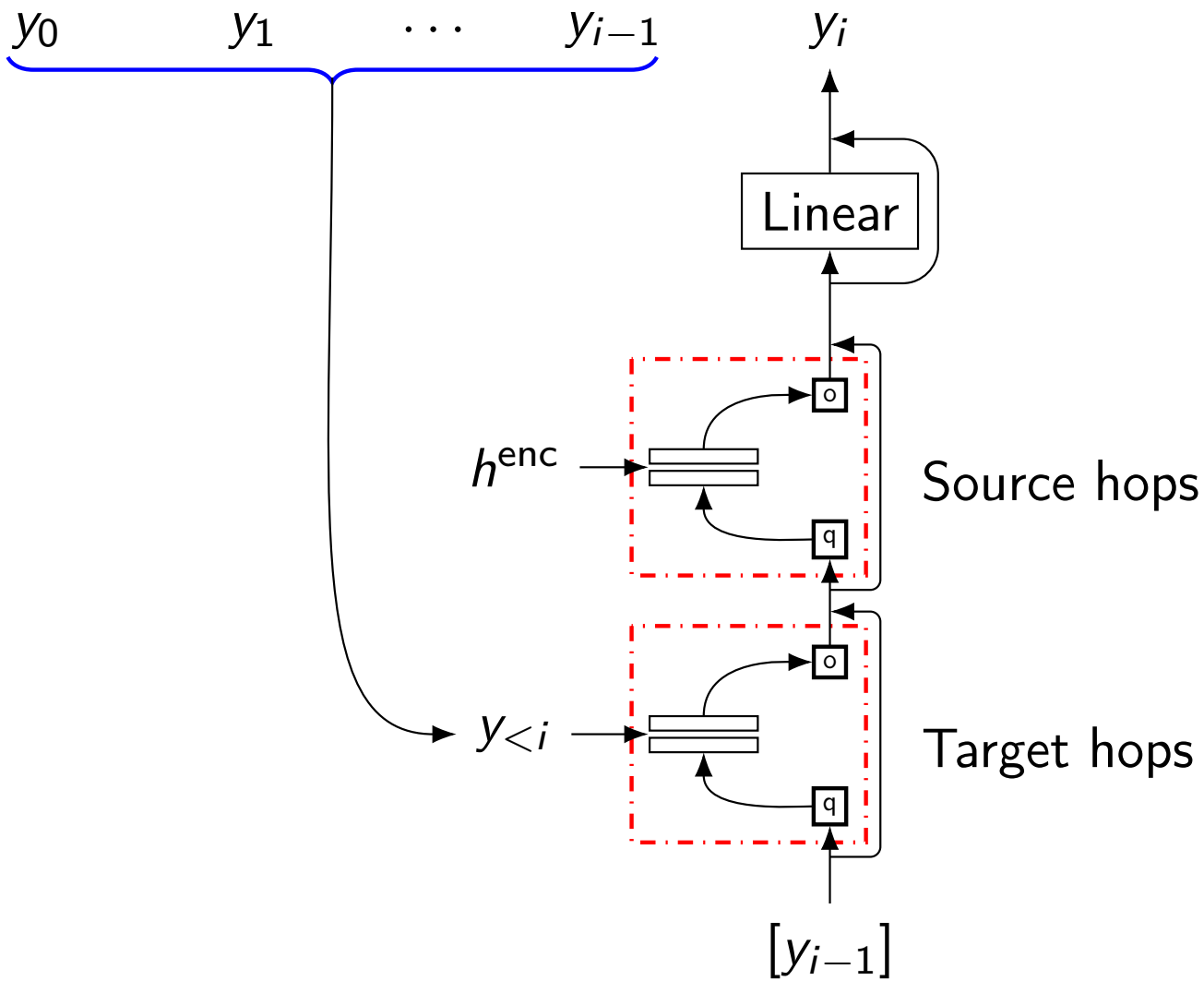
Self Attention



Removing the RNN



Their Decoder



Details

- Multi-head attention
- Position encoding:

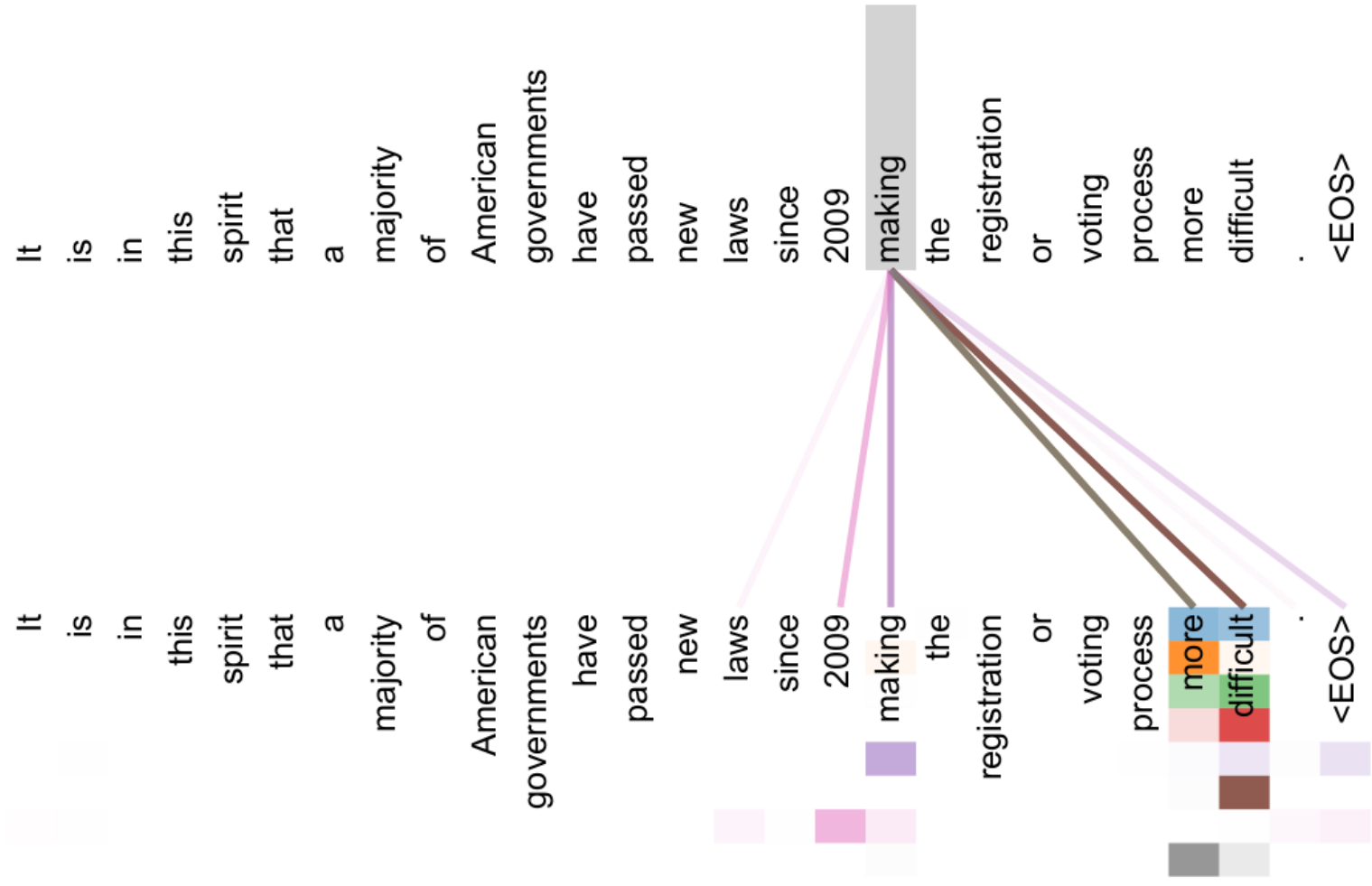
$$PE_{pos,2i} = \sin\left(pos/10000^{2i/d_{model}}\right)$$

- Dropout Scaled Dot-Product Attention:

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

- Label smoothing, Layer normalization, ...

Self-Attention Example



Results

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [17]	23.75			
Deep-Att + PosUnk [37]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [36]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [31]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [37]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [36]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	