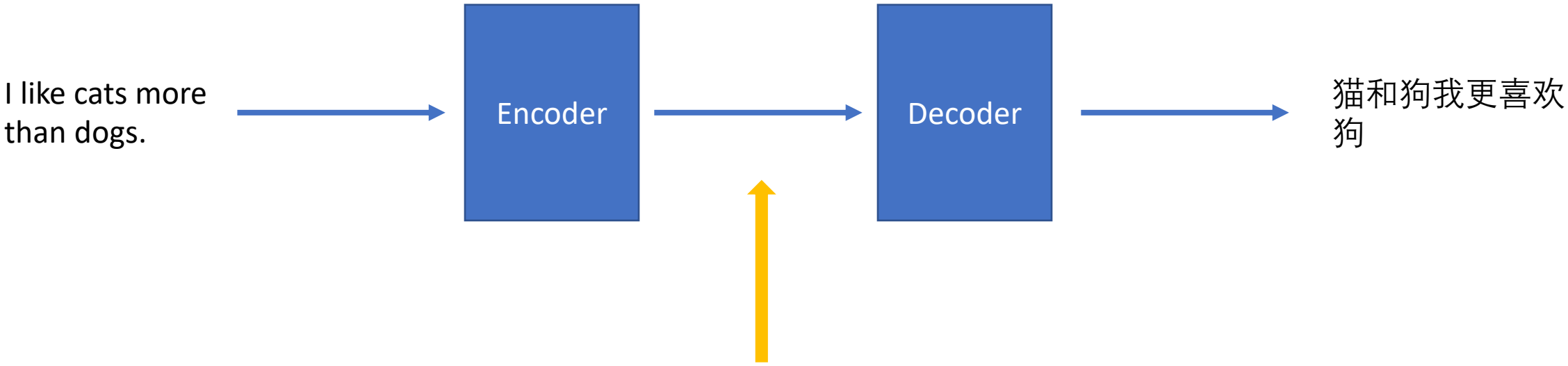
Attention Is All You Need

Presenter: Huidong Xie

Contributions

- Proposed network architecture (Transformer) reaches State-of-the-art result for machine translation. (English to German/ English to French)
- Training is highly parallelized. (RNN based architectures are hard to parallelize)
- Able to learn long-range dependencies within the input and output sequences. (RNN has poor long-range dependencies, RNN also has problems in short-range dependencies)
- Using simple explicit position encoding instead of learned position encoding.

Background on Neural Machine Translation



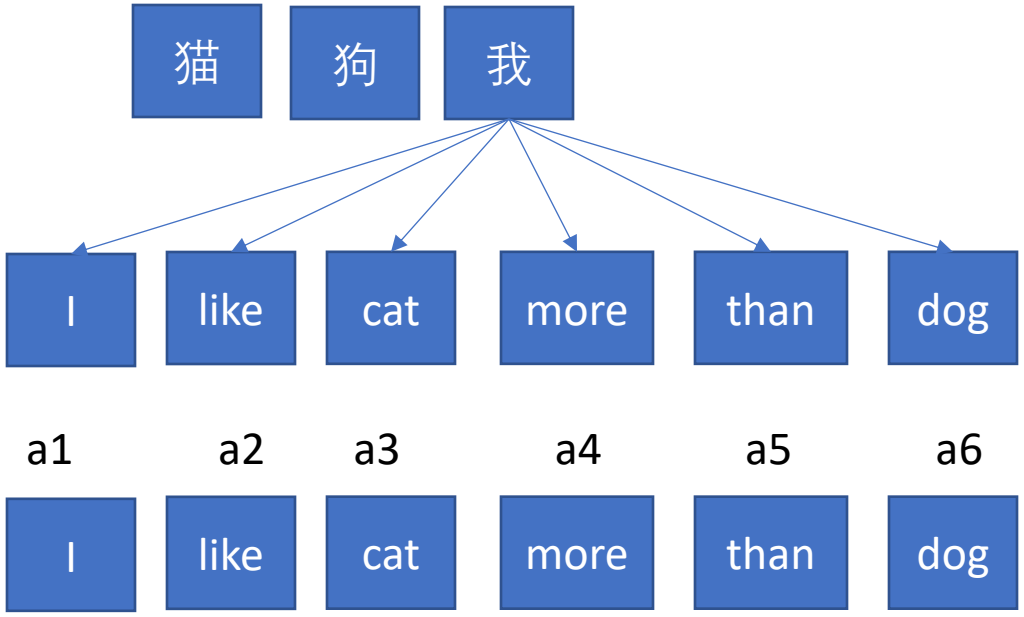
RNN Drawbacks

- Sequential nature of RNNs
- Difficulty of learning long-range(short-range) dependencies.
  - She is taller ***than*** me
  - I have no choice other ***than*** to write this paper
  - ***Neither*** he ***nor*** I knew about deep learning

Three Dependencies

- The input and output sequence
- The input sequence themselves
- The output sequence themselves

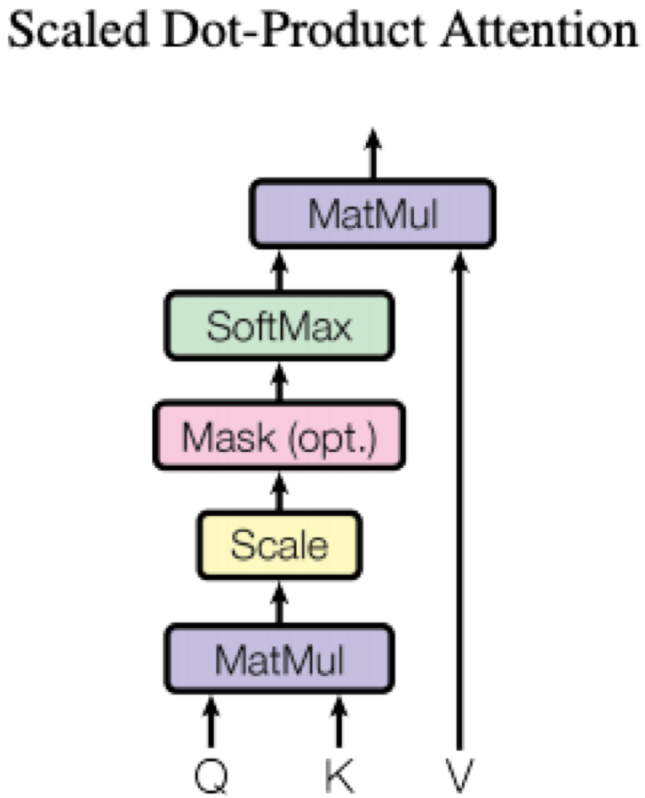
Single-head Attention mechanism



Single-head Attention mechanism

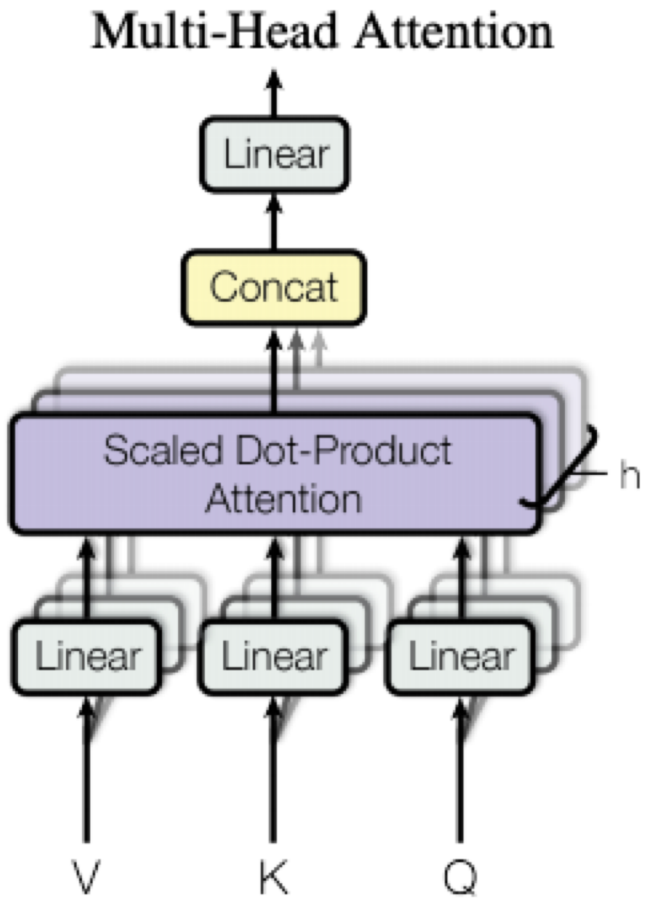
I like cats more than dogs

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

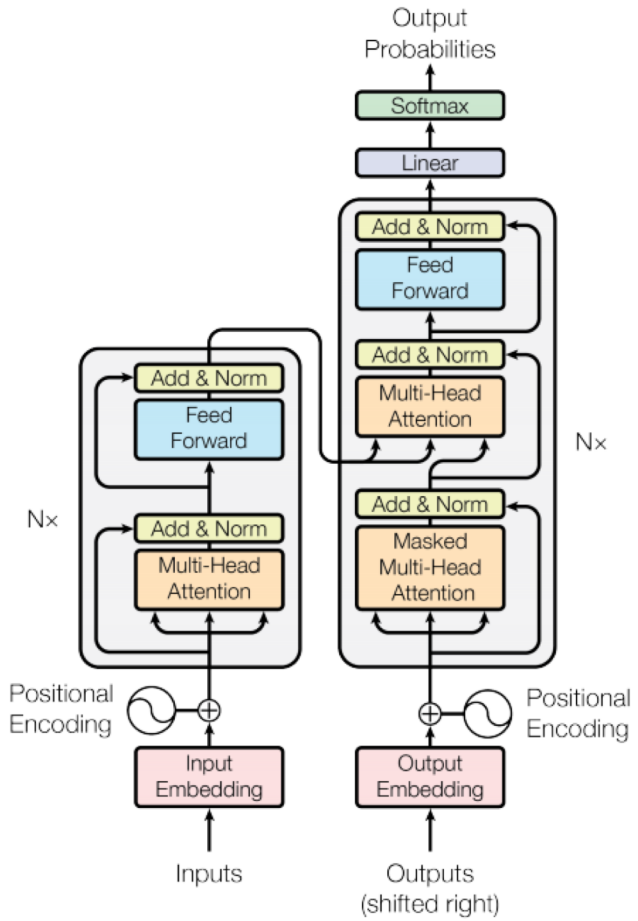


Multi-head Attention Mechanism

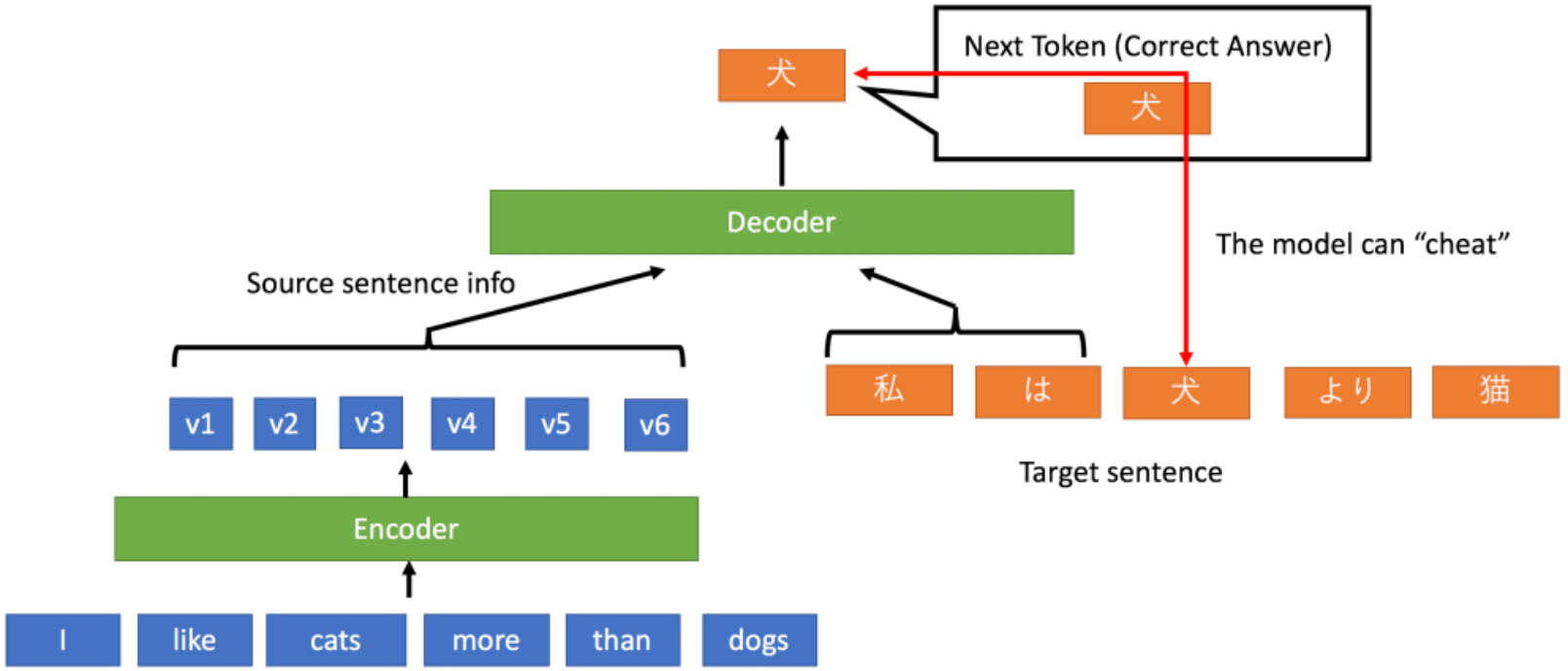
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Overall Model



Masked Multi-head Attention



Positional encoding

I like dogs more than cats

I like cats more than dogs

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

Results

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	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 [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	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**	**41.8**	$2.3 \cdot 10^{19}$	

Attenuation Visualization