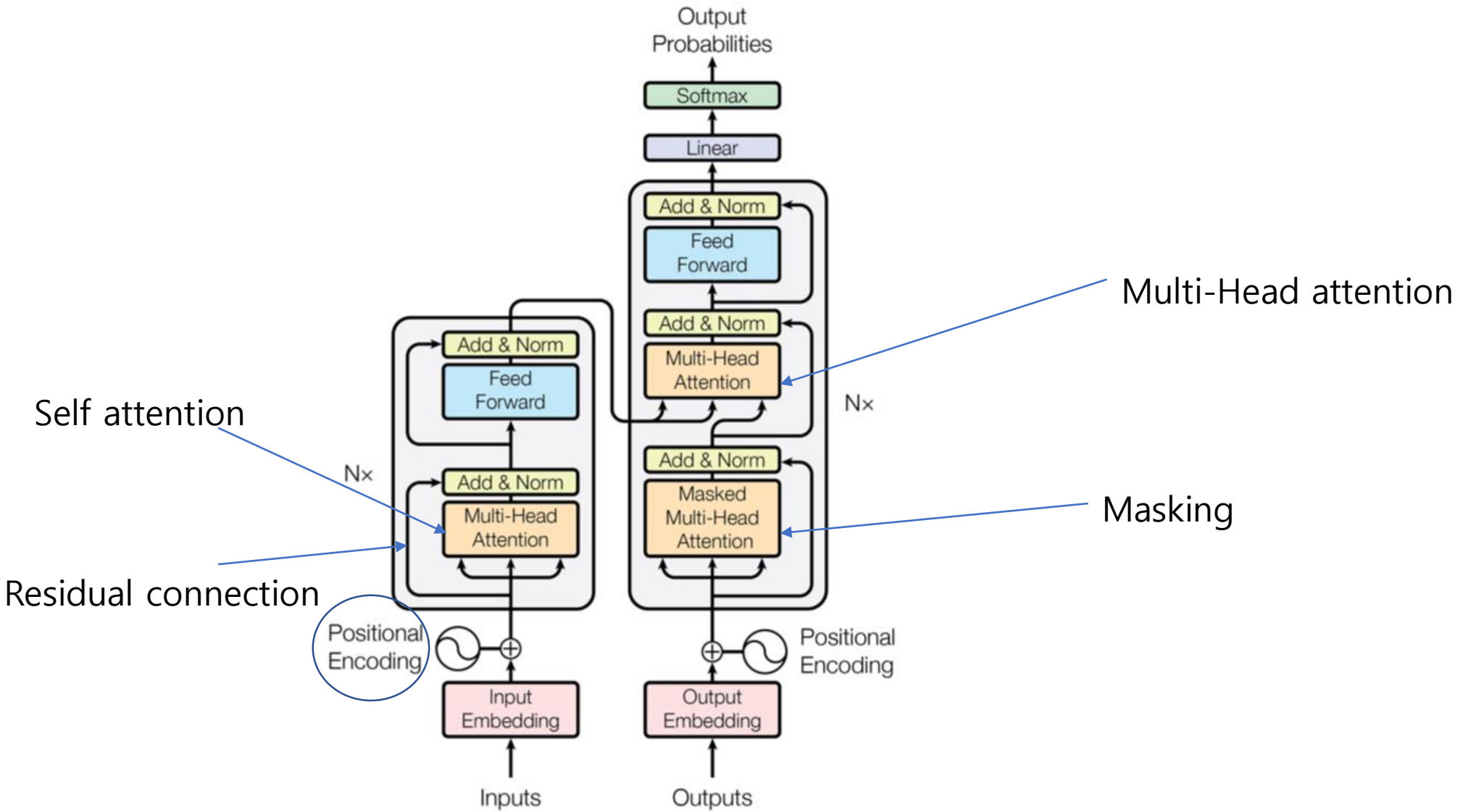
Semantics of the paper

Attention is all you need

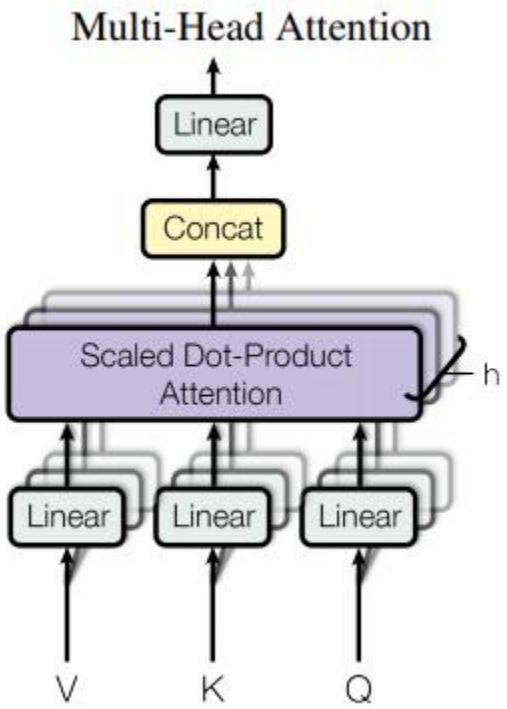
컴퓨터과학과 강효림

목차

Overall structure



Multi-Head attention



- Or multiple independently learned heads (Vaswani et al. 2017)

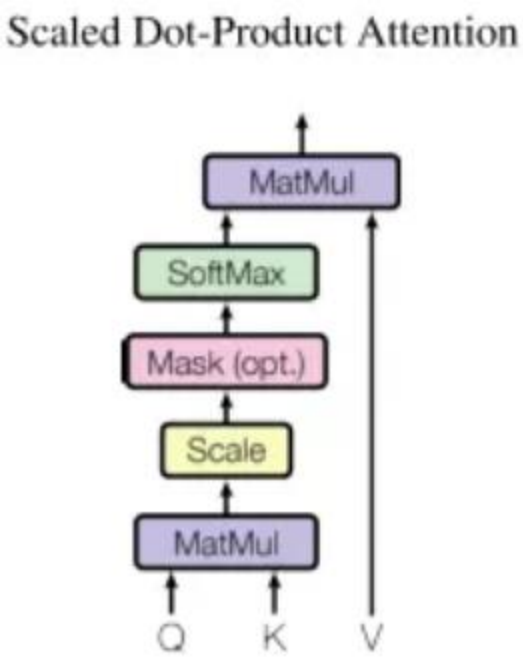
Multi-Head attention

- 취지
  - 복수의(plural) 단어 간 견련관계를 얻어내고 싶다.
- 방법론
  - 별도로 학습되는 linear transformation을 갖는 attention layers를 정의
  - Linear projection들은 q-k-v 간 서로 다른 관계를 추출하도록 학습됨
  - 각 attention layer의 output을 concat 후 linear projection

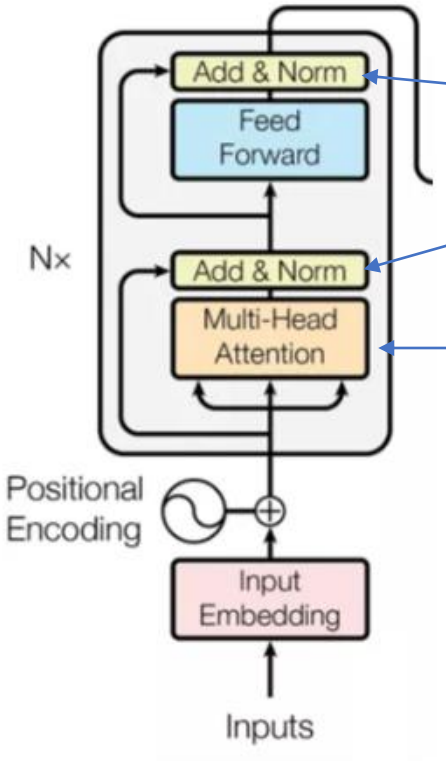
Scaled dot product attention

- 취지
  - Parameter 없는 attention 중 dot가 효율이 좋음
  - 그러나 naïve dot 연산은 dimension이 증가하면 값이 매우 커짐
  - 따라서 scaling 적용
- 식

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



Encoder



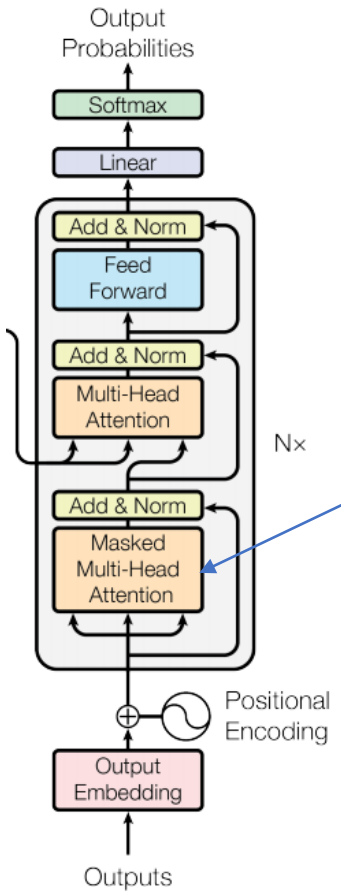
Layer normalization

q, k, v, 모두 자기 자신
-> self attention

Encoder

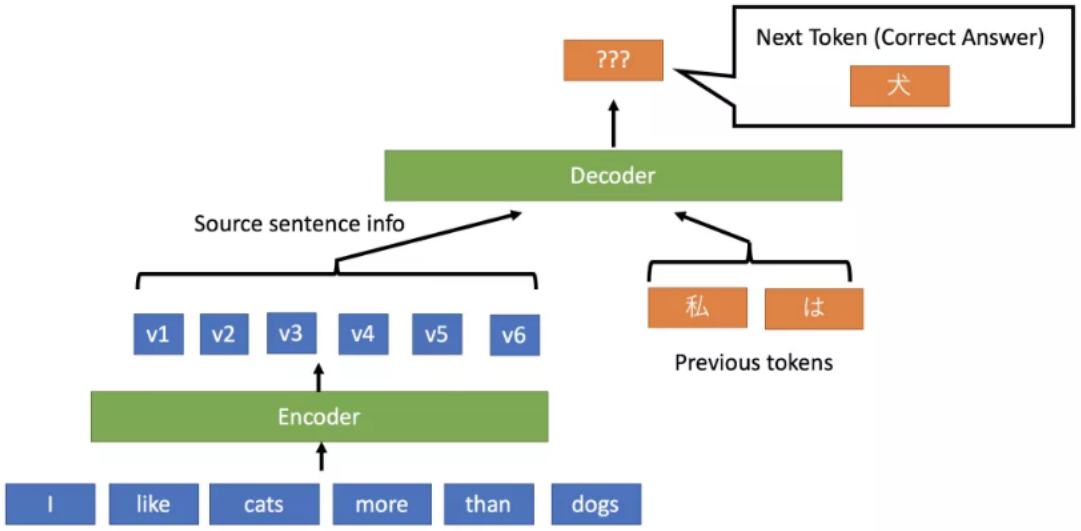
$$y = \mathrm{LayerNorm}(x + \mathrm{Sublayer}(x))$$

Decoder

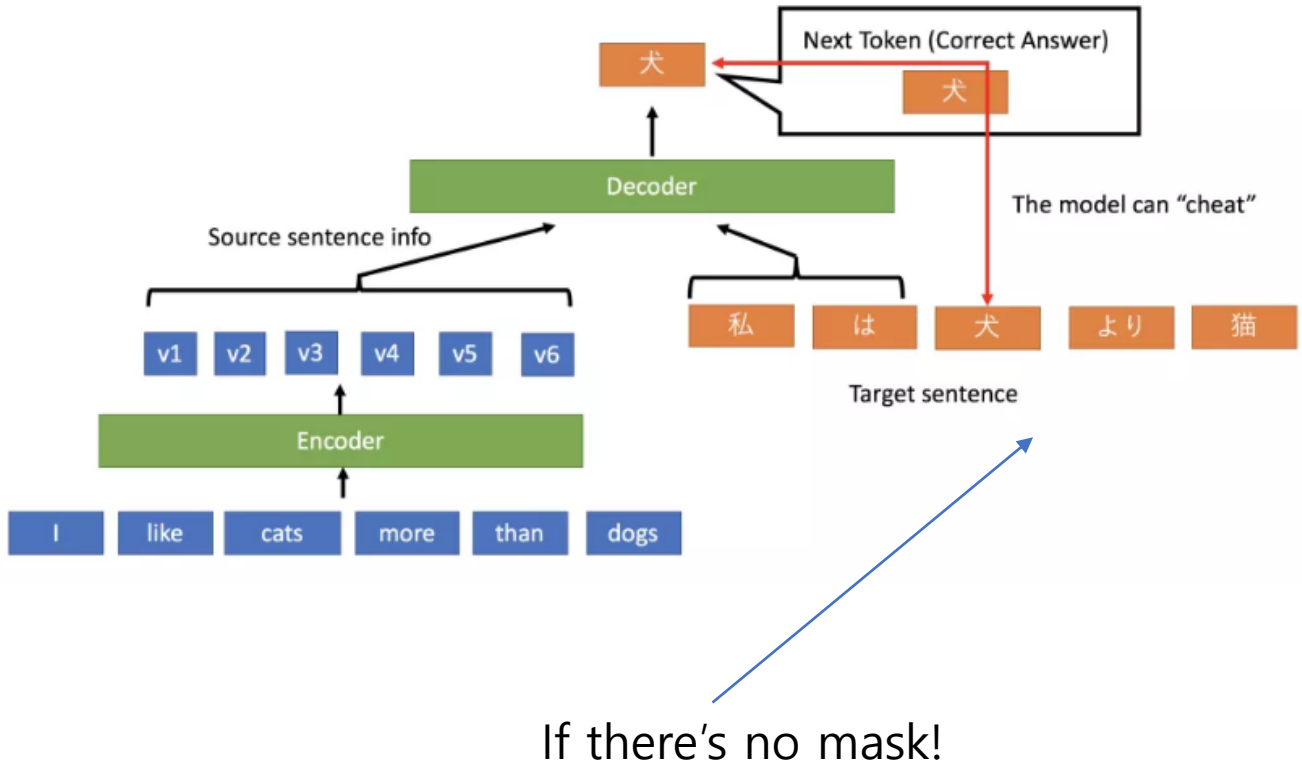


Masking?!

Masking - 私は犬よりも猫が好き



Correct way



If there’s no mask!

Positional Encoding

- 취지
  - RNN 사용하지 않기에, 단어의 위치 정보가 완전히 소실됨
  - Ex) I like cats more than dogs/I like dogs more than cats 가 동일하게 취급
  - 따라서 단어의 position 정보를 명시적으로 encoding해서 더해줌
- 식

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

Easy to learn relative position

Trivial facts/Experiments Result

- Optimizer는 Adam
- Residual Dropout 적용
- Label smoothing

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**	$2.3 \cdot 10^{19}$	

Thank you for listening

Let’s talk about it more!