Attention Is All You Need

https://arxiv.org/abs/1706.03762

0. Introduction

- 기존 시퀀스 모델(RNN, LSTM 등)은 순차적 처리 방식이라 병렬화가 어렵고, 긴 문장일수록 정보 전달이 약해짐.
- CNN 기반 모델은 병렬 처리에 강하지만, 긴 거리 의존성(long-range dependency)을 잡기 위해 깊은 구조가 필요함.
- 이 논문은 순수 Attention 메커니즘만으로 시퀀스를 처리하는 Transformer 구조를 제안함.
- 목적: 빠른 학습, 병렬성 확보, 장기 의존성 처리 개선.

1. Overview

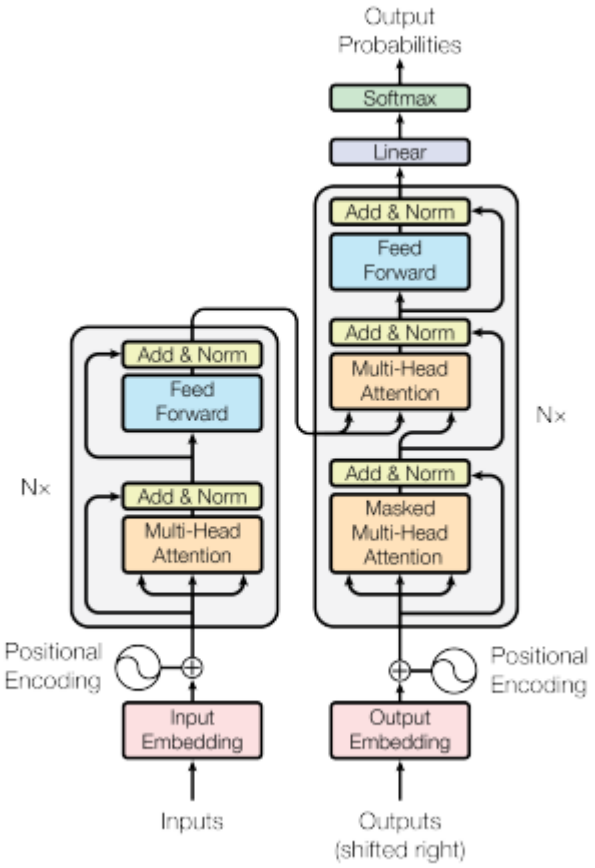
- Transformer는 encoder-decoder 구조를 따름
- encoder는 입력 시퀀스를 정제된 표현으로 변환하고, decoder는 이 표현을 기반으로 출력 시퀀스를 생성함
- 각 블록은 다음과 같은 공통 구조로 이루어짐: multi-head attention, position-wise feed-forward network, residual connection + layer normalization
- 입력 순서를 인식하기 위해 positional encoding이 추가됨
- 모든 연산이 병렬로 처리되어 속도와 효율이 크게 개선됨

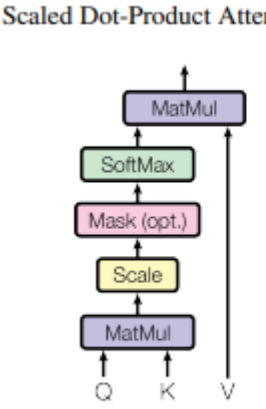
2. Challenges

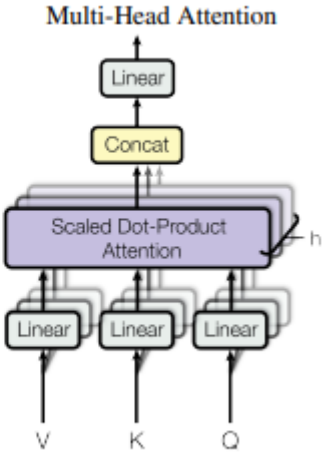
- RNN 계열 모델은 입력을 순차적으로 처리해야 해서 학습 속도가 느리고 병렬화가 어렵다
- CNN 계열 모델은 병렬화는 가능하지만, 문맥의 길이가 늘어나면 receptive field를 키워야 해서 구조가 복잡해짐
- 두 방식 모두 긴 거리의 토큰 간 관계 학습에는 한계가 존재함

- 이를 극복할 수 있는 새로운 구조가 필요했고, Transformer는 그 대안으로 제시됨

3. Method







- scaled dot-product attention

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

  - query, key, value 간 유사도를 계산하고, 그에 따라 정보를 가중합함
  - 연산 안정성과 스케일 조절을 위해 score를 sqrt(d_k)로 나눔
- multi-head attention

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

  - 여러 attention head를 병렬로 사용해 다양한 의미 공간에서 정보를 추출함
  - 문법적, 의미적 다양한 관계를 동시에 학습 가능
- position-wise feed-forward network

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

  - 각 위치별로 동일한 MLP를 적용하여 정보를 정제하고 비선형성을 부여함
  - attention 결과에 대한 추가 표현 학습 수행
- positional encoding

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

  - attention 구조는 순서 정보가 없기 때문에, 위치 정보를 인코딩해서 입력에 추가함
  - 사인/코사인 기반의 주기 함수로 위치를 표현함
- residual connection + layer normalization
  - 각 블록의 입력을 출력에 더해 잔차 연결을 수행하고, 정규화로 학습을 안정시킴
  - 깊은 구조에서도 정보 소실 없이 학습 가능하게 만듦
- encoder-decoder 구조
  - encoder는 self-attention과 feed-forward block을 N번 반복

- decoder는 masked self-attention, encoder-decoder attention, feed-forward block 순으로 구성됨
- decoder의 마스킹은 미래 단어를 보지 못하도록 제한하기 위함

4. Experiments

- 데이터셋: WMT 2014 English-German, English-French
- 모델 구성
  - base: 6-layer encoder & decoder, hidden size 512, attention heads 8
  - big: hidden size 1024, attention heads 16
- 학습 설정
  - optimizer: Adam
  - learning rate: warm-up 후 decay
  - regularization: dropout, label smoothing 사용

5. 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**	$2.3 \cdot 10^{19}$	

Table 3: Variations on the Transformer architecture. Unlisted values are identical to those of the base model. All metrics are on the English-to-German translation development set, newstest2013. Listed perplexities are per-wordpiece, according to our byte-pair encoding, and should not be compared to per-word perplexities.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser el al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser el al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	93.3

- BLEU 점수 기준 기존 RNN/CNN 기반 모델보다 우수한 성능을 기록함
- English-German: base 모델이 기존보다 빠른 수렴과 더 나은 정확도
- English-French: big 모델에서 최고 성능 달성
- 학습 시간 단축과 병렬성 확보로 실용성 높음

6. Insight

- attention만으로도 시퀀스 처리와 장기 의존성 학습이 가능함을 입증
- 이후 BERT, GPT 등 주요 언어 모델의 기반이 됨
- 계산량이 $O(n^2)$이라 긴 시퀀스 처리에서 비효율적일 수 있음
- 이 문제를 해결하기 위해 이후 다양한 경량화/개선 구조들이 등장함 (예: Linformer, Performer 등)