Attention is all you need (2017)

태그 Transformer

세상을 바꾼 논문: "Attention Is All You Need" 집중 탐구

자연어 처리(NLP) 분야를 공부하거나 관련 연구를 진행한다면, "Attention Is All You Need"라는 제목의 논문을 한 번쯤은 들어보셨을 겁니다. 2017년 Google Brain 팀이 발표한 이 논문은 단순히 새로운 모델을 제시한 것을 넘어, NLP 연구의 패러다임을 바꾸고 이후 등장하는 수많은 혁신적인 모델들의 초석이 되었습니다. 오늘은 바로 이 기념비적인 논문, **트랜스포머(Transformer)**를 세상에 알린 "Attention Is All You Need"를 깊이 있게 리뷰하며 그 핵심 아이디어와 영향력을 되짚어보겠습니다.

---

1. Attention Is All You Need

이 논문 이전까지, 기계 번역과 같은 시퀀스-투-시퀀스(Sequence-to-Sequence) 작업은 주로 순환 신경망(RNN)이나 합성곱 신경망(CNN)에 기반한 모델들이 주를 이루었습니다. 하지만 이 모델들은 긴 시퀀스 처리의 어려움, 순차적 계산으로 인한 병렬화의 한계 등의 문제를 안고 있었습니다. "Attention Is All You Need"는 이러한 관습을 과감히 깨고, 오직 **어텐션(Attention) 메커니즘**만을 사용하여 기존 모델들의 성능을 뛰어넘는 새로운 아키텍처, **트랜스포머**를 제안했습니다. 이 논문은 단순히 번역 품질을 향상시킨 것을 넘어, 모델 학습의 효율성을 극대화하고, 이후 BERT, GPT와 같은 대형 언어 모델(LLM) 탄생의 직접적인 계기가 되었다는 점에서 그 중요성이 매우 큽니다.

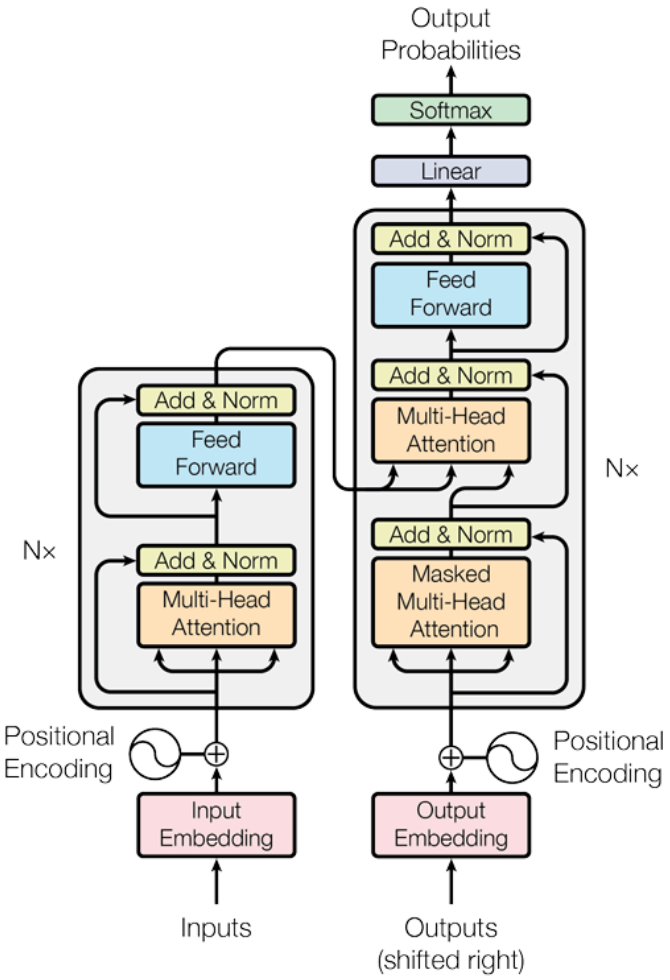
2. 기존의 한계를 넘어서: 왜 '어텐션'인가?

기존의 대표적인 시퀀스 모델인 RNN(특히 LSTM, GRU)은 단어와 같은 시퀀스 데이터를 순차적으로 처리합니다. 이는 각 시점의 계산이 이전 시점의 계산 결과에 의존하는 구조로, 문장이 길어질수록 정보가 소실되거나(장거리 의존성 문제) 계산이 느려지는 단점이 있었습니다.

어텐션 메커니즘은 이러한 한계를 극복하기 위한 아이디어로, 쉽게 말해 "중요한 부분에 집중하자"는 개념입니다. 번역을 예로 들면, 출력 단어를 생성할 때 입력 문장의 특정 단어들에 더 많은 가중치를 부여하여 참고하는 방식입니다. 이 논문은 여기서 한 걸음 더 나아가, "어텐션 메커니즘만으로도 충분히 뛰어난 모델을 만들 수 있지 않을까?"라는 혁신적인 질문을 던졌고, 그 해답으로 트랜스포머를 제시했습니다.

3. 트랜스포머 아키텍처 완전 정복

트랜스포머는 대부분의 신경망 시퀀스 변환 모델처럼 **인코더-디코더(Encoder-Decoder)** 구조를 따릅니다. 입력 문장을 인코더가 처리하여 문맥 정보를 담은 표현으로 만들고, 디코더는 이 표현을 바탕으로 출력 문장을 생성합니다.



3.1. 전체 그림: 인코더-디코더 구조

- **인코더(Encoder)**: 입력 시퀀스(예: 한국어 문장)를 받아 각 단어의 의미와 문맥 정보를 함축한 표현(벡터 시퀀스)으로 변환합니다. 여러 개의 동일한 인코더 레이어를 쌓아 구성됩니다.
- **디코더(Decoder)**: 인코더의 출력과 이전에 생성된 출력 단어들을 입력으로 받아 다음 단어를 예측합니다. 역시 여러 개의 동일한 디코더 레이어를 쌓아 구성됩니다.

3.2. 핵심 엔진: 어텐션 메커니즘 파헤치기

트랜스포머의 심장이라고 할 수 있는 어텐션 메커니즘은 주로 **스케일드 닷-프로덕트 어텐션(Scaled Dot-Product Attention)**을 사용합니다.

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

여기서 Q (Query), K (Key), V (Value)는 입력 벡터로부터 파생된 벡터들입니다. 쿼리가 특정 키와 얼마나 유사한지를 계산하고(내적), 그 유사도를 정규화(소프트맥스)하여 가중치를 얻은 뒤, 이 가중치를 값 벡터에 곱해 최종 어텐션 값을 얻습니다. dk로 나누는 것은 내적 값이 너무 커지는 것을 방지하여 학습 안정성을 높이기 위함입니다.

멀티-헤드 어텐션 (Multi-Head Attention)은 이러한 어텐션을 여러 개 병렬로 수행하는 방식입니다. 마치 여러 사람이 각기 다른 관점에서 정보를 살펴보는 것과 같습니다. 각 "헤드"는 서로 다른 표현 부분 공간(representation subspace)에서 정보를 학습하고, 이 결과들을 종합하여 더 풍부한 정보를 얻습니다. 논문에서는 8개의 헤드를 사용했습니다.

트랜스포머는 이 멀티-헤드 어텐션을 세 가지 다른 방식으로 활용합니다:

1. **인코더 셀프 어텐션 (Encoder Self-Attention)**: 인코더 내부에서 입력 문장 내 단어들 간의 관계를 파악합니다. 예를 들어, "그것"이라는 대명사가 문장 내 어떤 명사를 가리키는지 등을 학습할 수 있습니다.
2. **디코더 마스크드 셀프 어텐션 (Decoder Masked Self-Attention)**: 디코더 내부에서 이미 생성된 출력 단어들 간의 관계를 파악합니다. 단, 미래의 단어를 미리 볼 수 없도록 마스킹(masking) 처리를 합니다. 이는 번역과 같이 순차적으로 결과를 생성해야 하는 자기 회귀(auto-regressive) 특성을 유지하기 위함입니다.
3. **인코더-디코더 어텐션 (Encoder-Decoder Attention)**: 디코더가 출력 단어를 생성할 때, 인코더가 처리한 입력 문장의 어떤 부분에 주목해야 할지를 결정합니다. 이것이 번역 과정의 핵심적인 연결고리 역할을 합니다.

3.3. 인코더 상세 분석

각 인코더 레이어는 두 개의 주요 하위 레이어로 구성됩니다:

- **멀티-헤드 셀프 어텐션 레이어**: 위에서 설명한 것처럼 입력 시퀀스 내의 관계를 파악합니다.
- **위치별 피드포워드 네트워크 (Position-wise Feed-Forward Network)**: 어텐션 레이어를 통과한 각 위치의 표현을 독립적으로 변환하는 간단한 완전 연결 신경망입니다. 모든 위치에서 동일한 가중치를 공유하지만, 각 위치별로 적용됩니다.

각 하위 레이어 주위에는 **잔차 연결(Residual Connection)**과 **레이어 정규화(Layer Normalization)**가 적용됩니다. 이는 깊은 네트워크에서 학습을 안정화하고 정보 흐름을 원활하게 하는 데 도움을 줍니다.

3.4. 디코더 상세 분석

각 디코더 레이어는 인코더의 두 하위 레이어 외에 한 가지 하위 레이어를 더 가집니다:

- **마스크드 멀티-헤드 셀프 어텐션 레이어**: 디코더의 이전 출력들에 대한 셀프 어텐션입니다.
- **인코더-디코더 어텐션 레이어**: 인코더의 전체 출력 시퀀스에 대해 어텐션을 수행합니다.
- **위치별 피드포워드 네트워크**: 인코더와 동일합니다.

마찬가지로 잔차 연결과 레이어 정규화가 각 하위 레이어에 적용됩니다.

3.5. 순서 정보를 기억하는 방법: 위치 인코딩 (Positional Encoding)

트랜스포머는 RNN과 달리 순차적인 처리 구조가 없기 때문에, 단어의 순서 정보를 모델에 알려줄 별도의 방법이 필요합니다. 이를 위해 각 단어의 임베딩 벡터에 **위치 인코딩 (Positional Encoding)** 벡터를 더해줍니다. 이 논문에서는 고정된 사인(sine) 및 코사인(cosine) 함수를 사용하여 위치 정보를 표현했습니다. 이를 통해 모델은 단어의 절대적인 위치와 상대적인 위치 관계를 학습할 수 있습니다.

4. 놀라운 성능과 그 의미

트랜스포머는 주요 기계 번역 벤치마크인 WMT 2014 영어-독일어, 영어-프랑스어 번역 작업에서 당시 최고 성능(State-of-the-Art, SOTA)을 달성했습니다. 특히, 기존 모델들보다 훨씬 적은 학습 시간으로 더 높은 번역 품질(BLEU 점수)을 보여주었습니다.

- **WMT 2014 영어-독일어**: 28.4 BLEU (이전 SOTA보다 2.0 이상 높음)
- **WMT 2014 영어-프랑스어**: 41.8 BLEU (단일 모델 최고 성능)

이러한 성능 향상 외에도 트랜스포머의 진정한 의미는 다음과 같습니다:

- **뛰어난 병렬 처리**: 순환 구조가 없어 모든 단어에 대한 계산을 병렬로 처리할 수 있어 학습 속도가 매우 빠릅니다. 이는 GPU 활용을 극대화합니다.
- **장거리 의존성 효과적 학습**: 어텐션 메커니즘은 문장 내 멀리 떨어진 단어 간의 관계도 직접적으로 파악할 수 있어 장거리 의존성 문제에 강합니다.
- **다른 태스크로의 일반화**: 기계 번역뿐만 아니라 영어 구문 분석(Constituency Parsing)과 같은 다른 NLP 태스크에서도 좋은 성능을 보여 범용적인 아키텍처로서의 가능성을 입증했습니다.

5. 트랜스포머가 남긴 것들: 영향력과 미래 전망

"Attention Is All You Need"가 발표된 이후, 트랜스포머 아키텍처는 NLP 분야의 표준으로 자리 잡았습니다.

- **후속 모델의 기반**: BERT, GPT, T5, XLNet 등 우리가 현재 접하는 대부분의 고성능 대형 언어 모델(LLM)은 트랜스포머를 기반으로 하거나 그 변형을 사용합니다.

- **다양한 분야로의 확장**: NLP뿐만 아니라 컴퓨터 비전(Vision Transformer), 음성 인식, 신약 개발 등 다양한 분야로 트랜스포머의 아이디어가 확장되고 있습니다.

물론 트랜스포머에도 **한계점**은 존재합니다.

- **연산량**: 셀프 어텐션은 시퀀스 길이에 제곱으로 비례하는 연산량($O(n2 \cdot d)$)을 가지므로, 매우 긴 시퀀스를 처리하는 데는 계산 비용이 많이 듭니다. 이를 개선하기 위한 경량화된 어텐션(Sparse Attention, Linformer, Reformer 등) 연구가 활발히 진행 중입니다.
- **위치 정보**: 고정된 위치 인코딩이나 학습 가능한 위치 임베딩이 사용되지만, 순서 정보를 처리하는 방식에 대한 연구는 계속되고 있습니다.

그럼에도 불구하고, 트랜스포머는 어텐션이라는 비교적 간단한 아이디어를 통해 복잡한 시퀀스 데이터를 효과적으로 모델링할 수 있음을 보여주었고, 이는 딥러닝 연구에 큰 영감을 주었습니다.

6. 맺음말: 어텐션, 모든 것의 시작

"Attention Is All You Need"는 논문 제목처럼 '어텐션'이라는 핵심 아이디어로 NLP 분야에 혁명적인 변화를 가져왔습니다. 순환이나 합성곱 없이도 뛰어난 성능을 달성할 수 있음을 증명하며, 이후 수많은 연구의 방향을 제시했습니다. 이 논문을 이해하는 것은 현대 AI, 특히 자연어 처리 기술의 근간을 이해하는 첫걸음이라고 할 수 있습니다. 트랜스포머의 등장으로 우리는 더욱 정교하고 인간과 유사한 수준의 언어 이해 및 생성이 가능한 AI 시대로 한 걸음 더 다가섰습니다. 앞으로 이 강력한 아키텍처가 또 어떤 놀라운 발전을 이끌어낼지 기대됩니다.