**자연어(딥러닝) 연구조사**

Attention is all you need

Liu, Pengfei, et al. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *ACM Computing Surveys* 55.9 (2023): 1-35.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

과목명 : 소프트웨워응용

학과 : 컴퓨터융합소프트웨어학과

학번 : 2021270678

이름 : 나강민

"Attention is All You Need" 논문에서 제안된 Transformer 모델은 다양한 자연어 처리 작업에 적용할 수 있습니다. 기본적으로 Transformer은 입력 시퀀스를 출력 시퀀스로 변환하는 Sequence-to-Sequence 작업을 수행합니다. 요약, 질문-답변, 문장 생성과 같은 다양한 자연어 처리 작업에 사용할 수 있습니다.

기존에는 RNN과 CNN을 기반으로 인코더-디코더를 이용했지만, 해당 논문에서는 recurrence와 convolution을 제외하고 오직 attention mechanism에만 기반한 Transformer라는 간단한 모델을 제안합니다. 이는 데이터를 병렬적으로 처리하기 때문에 재귀적으로 처리하던 고질적인 문제를 효과적으로 해결하며, 더 빠른 속도로 처리할 수 있습니다.

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명위의 그림은 Transformer의 모델 구조를 보여줍니다.왼쪽은 인코더 구조이고, 오른쪽은 디코더의 구조입니다. Input 값을 Input Embedding을 통해 행렬값으로 변환합니다. 그리고 Positional Encoding을 통해 Input vector값의 문장 위치에 대한 정보를 넣어주게 됩니다. 이후 Multi-Head Attention을 실행하게 되고 Input Embedding을 한 vector와 positional Encoding을 한 vector을 더한 값과 비교하여 normalization을 적용합니다. 이러한 Encoder의 동작과정에서 나온 최종 vector은 다시 Decoder에 Multi-head Attention layer으로 들어가 반복하여 더 정확한 결과값을 도출하여 최종 Output으로 나오게 됩니다.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명전체적인 flow는 이와 같으며 여기서 중요하게 봐야하는 부분은 Positional Encoding과 Multi-Head Attention이 무엇인지 살펴봐야 합니다.

해당 논문에서 ‘Scaled Dot-Product Attention’이란 Multi-Head Attention 내부에서 반복하는 Attention function입니다. 문장의 각 위치의 원소들 간의 상대적인 중요도를 계산을 하기위해 사용합니다.

폰트, 텍스트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명위 식은 Scaled Dot-Product Attention의 Output을 구하기 위한 공식입니다. Input으로 Query, Key, Value 세개가 들어오게 됩니다. Query 현재 처리되고 있는 단어의 Vector입니다. Key는 모든 문장들의 Stack 행렬입니다. Value는 Input Embedding값 입니다. 이와 같은 Vector들을 행렬곱 하는 과정인데, 현재 처리되고 있는 하나의 Query 단어인 vector와 모든 단어의 key vector를 곱하여 attention score를 구합니다. 수식으로는 로 나타내며 이 값은 기본적으로 vector의 값을 곱한 값이기 때문에 값이 비례해서 커집니다. 그래서 값을 나눠줍니다. 이후 softmax를 통해 Attention distribution을 구한 후 value vector와 곱한 이후 모든 값을 더하여 Attention Value vector가 완성됩니다. 위 Scaled Dot-Product Attention의 과정중 mask를 제외하고 전체적인 flow이며, mask는 문장에서 제외할 단어들을 설정할 수 있는 기능입니다.

텍스트, 스크린샷, 도표, 디자인이(가) 표시된 사진

자동 생성된 설명위와 같이 Attention 과정을 반복하여 사용하면 Multi-Head Attention이 됩니다. 여기서 Head는 사용자가 임의의로 설정가능한 값이며 이 값만큼 앞선 Scaled Dot-Product Attention을 반복합니다. 즉 다양한 관점에서 문장을 해석이 가능합니다. CNN과 같은 필터를 통해 convolution output을 만드는 방식으로 구현되어 있습니다. Linear 이후 h개로 구분된 각각의 query 쌍들을 만들어내는데, 이때 각기 다른 head값을 내보내어 출력과 입력 값의 차원이 같아지도록 한 다음 Concat을 거쳐 Linear으로 출력 값을 만듭니다

‘Attention is all you need’ 에서 중요한 개념인 Multi-Head Attention 이외에 Positional Encoding은 결국 문장의 관계를 구하는데 RNN은 순차적으로 처리해서 괜찮지만 Attention만 이용하는 이 모델은 어떻게 Embedding Vector의 위치는 어떻게 결정할 것인가? 라는 문제가 남아있습니다.

텍스트, 폰트, 화이트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명위 식은 방금 앞서 말한 문제를 해결하는 공식입니다. 인덱싱을 시키는 방법은 매우 많지만 현재 논문에서는 이러한 방법을 채택했습니다. 만약 "Hello World"라는 단어가 들어왔을 때, "Hello"의 위치가 (0, 2)일 때 pos=0, i=2를 뜻합니다. 즉 Embedding vector와 위 공식으로 계산한 Positional Encoding Vector값이 더해진 결과값을 이용하여 모델구조를 따라가게 됩니다.

즉 Encoder는 위에 설명을 통해 과정이 동작되고 동작된 결과값의 Query, Key값이 Decoder에 Attention으로 들어가게 되며 Attention layer를 들어가기 전 학습과정에 Masked-Multi-Head Attention을 통해 Label을 알고 있는 상태에서 학습을 방지하고자 해당 원소만 비교하기 위해 나머지 위치는 mask로 가려 학습을 진행시킵니다.

해당 논문의 Contribution은 기존의 RNN이나 CNN과는 달리 Self-Attention 메커니즘을 사용하여 입력 시퀀스 내의 모든 위치를 동시에 처리함으로써 병렬 처리가 가능해지고 학습 속도가 빨라지며, 입력 시퀀스 내의 상호작용을 더 잘 모델링할 수 있다는 것입니다. 또한 기존 자연어 모델들은 입력 시퀀스의 길이가 길어지면 정보의 손실과 연산량의 증가로 인해 성능이 하락하였으나, Transformer는 이러한 문제를 해결하여 모든 자연어 처리에서 대안이 될 수 있다는 점입니다. 현재 다양한 자연어 처리 모델이 사용되고 있습니다. 기계 번역, 언어 모델링, 챗봇, 문서 요약 등에 사용됩니다. Chat GPT, PDF GPT, 구글 번역 등의 GPT 모델은 현재 전반적으로 자연어 처리에 사용되고 있습니다.