14. 주의

순환신경망은 정보를 내부 상태에서 순환시키는 방식으로 기억을 유지한다. 순환신경망에서 LSTM 등은 아주 멀리까지 정보를 전달할 수 있지만 그럼에도 여러 한계가 있다. 기계 번역에서 영어 "I am a student"를 한국어로 번역한다면 "나는.."으로 시작하게 될 것이다. 그런데 이 단계에서는 "I"의 정보면 충분하지 나머지 "am a student"의 정보는 별로 필요하지 않다. 그런데 LSTM의 내부 상태에는 출발 언어의 모든 문장의 정보가 담겨있기 때문에 불필요한 정보가 많아지게 된다.

이러한 문제를 해결하기 위해 도입된 것이 주의(attention)이다. 우리는 볼 때 눈 앞에 펼쳐진 모든 광경을 받아들이는 것이 아니라 그 중 특정한 부분에만 초점을 맞춰 받아들인다. 이것을 주의라고 한다. 주의는 처리해야 할 정보의 양을 줄여준다는 장점이 있다. 이러한 생물의 주의를 모방하여 이전의 모든 정보를 한꺼번에 전달하는 것이 아니라 그 중에 일부의 정보에만 초점을 맞추는 것인 신경망에서 주의이다. 물론 생물에서 주의는 지각(perception) 단계에서 이뤄지는 것인 반면에, 인공신경망에서 주의는 이전 단계의 정보에 대한 것으로 차이가 있다.

14.1. Seq2Seq에서 주의 메커니즘

Seq2Seq 모형은 인코더와 디코더로 이뤄져 있다. 인코더는 토큰 x_i를 입력으로 받고, 내부 상태 h_i를 다음 단계로 전달한다. 디코더는 이전 상태 s_{j-1}와 이전의 토큰 y_{j-1}을 입력 받아 토큰 y_j를 출력한다.

주의 메커니즘은 예를 들어 디코더에서 토큰 y_5와 관련된 부분이 인코더에서 h_3이라고 하면 h_3을 y_5에 직접 입력할 수 있게 만드는 것이다. 이를 위해 다음과 같은 계산 과정을 거친다.

먼저 디코더의 현재 상태 s_j과 인코더의 모든 상태 h_i의 유사성 e_i를 계산한다. 유사도를 계산하는 방법은 코사인 유사도와 비슷하지만 좀 더 복잡하다.

이렇게 구한 유사성들을 소프트맥스 함수에 넣어주면 주의 가중치 a_i를 구한다.

$$a_i = \text{SOFTMAX}(e_1, e_2, \cdots, e_n)$$

인코더의 각 상태에 주의 가중치를 곱하여 더한다. 이를 맥락(context)이라고 한다.

$$c_j = \sum_i a_i h_i$$

이제 이전 상태 s_{j-1}, 이전 토큰 y_{j-1}과 함께 맥락 c_j를 함께 입력으로 하여 y_j를 예측한다.

14.2. 트랜스포머

2017년 구글의 바스와니(Vaswani) 등이 발표한 트랜스포머 모형은 아예 순환신경망을 제거하고 주의 메커니즘만으로 시퀀셜 데이터를 다룰 수 있음을 보여주 었다.

먼저 토큰들이 $x_1, x_2, \cdots, x_n$이 있고 토큰 x_i를 입력으로 처리할 차례라고 해보자. 순환신경망의 경우에는 이전 상태 h_{i-1}을 넘겨 받지만 트랜스포머는 이전 토큰들 중에 현재 토큰과 관련이 깊은 것에 주의를 준다. 이를 위해서 토큰 x_i를 질의 벡터 q_i로 바꾼다. 다음으로 x_i 이전의 모든 토큰 x_j (단, $j = 1, 2, \cdots, i-1$)를 키 벡터 k_j와 값 벡터 v_j로 바꾼다. 토큰을 각 종류의 벡터로 바꾸는 방법은 임베딩과 마찬가지로 학습된다.

다음은 주의 메커니즘과 같이 q_i와 k_j의 유사성 e_j을 계산하고, 주의 가중치 a_j를 구한다. 그리고 값 벡터에 주의 가중치를 곱하여 맥락을 구한다.

여기서 중요한 것은 주의 메커니즘이 인코더와 디코더 간에 적용되는 것이 아니라 하나의 시퀀스를 처리하는 모형 내에 적용되는 것이다. 이를 자기 주의(self attention)이라고 한다. 물론 트랜스포머로 인코더와 디코더를 만들어 Seq2Seq 형태의 모형을 만드는 것도 가능하다.

트랜스포머는 매우 복잡한 구조를 가지고 있지만 현재 자연어 처리에서 순환신경망을 능가하는 성능을 보여주고 있다.