트랜스포머는 어떻게 RNN을 이겼나?

이동재

20240530

Abstract

트랜스포머 (Transformer)는 현재 가장 많이 사용되는 모델 구조이다. 트랜스포머는 기존 모델들이 보여주던 성능을 압도적으로 뛰어넘으면서 다양한 작업에 사용되고 있다. 트랜스포머의 어떤 측면이 이를 가능케 했을까? 본 글에서는 트랜스포머가 RNN, CNN과 같은 기존 모델 구조를 밀어낼 수 있었던 이유에 대해 다룬다.

트랜스포머는 자연어와 이미지, 비디오 등 모든 종류의 입력에서 널리 사용되고 있다. '이미지는 CNN, 자연어는 RNN'이라는 공식을 깨부수고 양 분야에서 최고의 모델 구조로 올라선 것이다. 트랜스포머가 승리할 수 있었던 핵심적인 이유에 대해 알아보자.

트랜스포머가 기존 모델 구조를 이길 수 있었던 가장 큰 이유는 병렬성이다. RNN은 순차적으로 계산을 진행하기 때문에 병렬성이 떨어진다. 병렬성이 떨어진다는 말은 모델 크기를 키우는데 한계가 있다는 것이다. 트랜스포머의 모든 연산은 고도로 병렬화가 가능하다. 크기 법칙[1] (Scaling law)에 따르면, 모델의 크기와 학습 데이터의 양이 커질수록 모델의 성능이 좋아진다. 트랜스포머는 병렬화된 연산을 통해 대량의 학습 데이터를 거대한 모델에 학습시킬 수 있게 만들었다.

또 다른 이유는 장거리 의존성을 잘 학습할 수 있다는 점이다. RNN은 장거리 의존성을 포착하는 데 어려움이 있다. 이는 RNN이 순차적으로 계산을 진행하기 때문에 발생한다. 거리가 먼 단어의 정보는 연산을 거듭하면서 희석된다. 주변 값을 활용하는 CNN의 컨볼루션 연산은 근본적으로 장거리 의존성을 포착하는데 적합하지 않다. 트랜스포머는 자가 주의 연산 통해 장거리 의존성을 잘 학습할 수 있다.

트랜스포머가 등장한 지 7년이 지났다. 여전히 트랜스포머를 대체할 구조는 제시되지 않고 있다. 최근 Mamba[2]를 비롯하여 새로운 구조가 제시되고 있지만, 여전히 대세는 트랜스포머다. 트랜스포머를 넘어설 모델이 등장할지 그 귀추가 주목된다.

[1] Jared Kaplan et al. Scaling Laws for Neural Language Models

[2] Gu, A. et al. Mamba: Linear-Time Sequence Modeling with Selective State Spaces