정말 정말 정말 큰 모델은 좋다?

이동재

20241017

Abstract

과적합 (Overfitting) 문제는 기계 학습에서 가장 널리 알려진 문제 중 하나로, 모델이 지나치게 복잡하여 훈련 데이터를 암기해버린 상태를 말한다. 그러나, 신경망 기반의 현대 기계 학습 모델은 크면 클수록 좋다는 것이 정설이다. 왜 이러한 모순이 발생하는 것일까? 본 글에서는 이중 강하 (Double Descent) 현상을 소개하고, 이것에 대한 나의 생각을 공유한다.

과적합은 데이터에 비해 지나치게 복잡한 모델을 사용하여 훈련 데이터를 암기해버린 상태를 말한다. 우리가 모델에게 기대하는 것은 데이터에 대한 일반화된 법칙인데, 모델이 기억할 수 있는 양이 너무 많기 때문에 일반화된 법칙을 찾지 않고 쉬운 길인 암기를 택하는 것이다.

그러나, 어중간하게 암기하는 것을 넘어 훈련 손실 (Training Loss)를 0으로 만드는 수준으로 모델이 커지게 되면 다시 성능이 좋아진다고 한다. 이러한 현상을 이중 강하 연상이라고 하며, 지금의 초대형 모델들이 등장하게 된 배경 중 하나다.

놀랍게도 이러한 현상은 신경망을 넘어 KNN, SVM, 결정 나무 등 기존 기계학습 모델에도 나타난다. 즉, 이중 강하 현상은 현상은 신경망에만 국한된 것이 아니라, 데이터 기반의 기계학습에 모두 적용되는 일반적인 현상임을 알 수 있다.

나는 이러한 현상이 암기와 이해의 전환 과정에서 발생하는 현상이라고 추측한다. 모델이 작은 경우 데이터를 다 암기할 수 없으므로 데이터를 이해하여 일반화된 원리를 찾아야 한다. 반면, 모델이 커지는 경우 모든 데이터를 암기하고, 잡음까지도 암기할 수 있다. 만약 모델이 어중간한 크기라면, 데이터를 암기하는 것도 아니고 이해하는 것도 아닌 어중간한 학습을 하게 된다.

이중 강하라는 실험적 경험을 통해 거대한 모델들이 등장했지만, 과연 그들이 데이터를 진정으로 이해하고 있는 것일까? 인간과 달리 거대 언어 모델은 지금의 성능을 얻기 위해 지구 상의 거의 모든 데이터를 학습해야 한다. 이러한 학습의 비효율성을 생각했을 때 이중 강하라는 현상에 의존하여 성능을 높이는 것이 옳은 방향이라고 볼 수 있을까? 이쯤에서 우리가 올바른 방향으로 나아가고 있는지 다시 한 번 생각해보자.