컴퓨터학콜로키움 **(COSE-405) :: 4**주차 네이버**AI** 김진화 박사님

Deep Learning beyond Memorization, Adversarial Risks, and Bias

강의 : 2022-09-28 / 작성 날짜 : 2022-09-28

고려대학교 컴퓨터학과 2017320108

고재영

기술이 급격하게 발전하고 우리 삶에 다양한 부분에 전통적으로 인간이 하던 일을 기계가 대체하는 영역이 보다 더 넓어짐에 따라 AI와 머신러닝에 대한 관심은 점진적으로 높아져왔다. CCTV나 자율주행 자동차와 같이 객체 탐지에 관련한 응용 뿐만 아니라 질의 응답 시스템, 문서 분류, 기계 번역과 같은 텍스트 및 음성 데이터와도 긴밀하게 연결되어 인공 신경망으로 딥러닝 기술의 중요도가 대두되었다. 이번 시간에는 네이버 AI에서 근무 중이신 동시에 서울대 AIIS 인공지능 대학원 겸임 교수이신 김진화 박사님께서 딥러닝에 관련한 강연을 하셨다.

이번 강의는 딥러닝의 정의에 대한 간단한 소개를 시작으로, 딥러닝을 연구하는 영역에 있어서 가장 큰 세 가지의 당면하는 과제에 초점을 맞추셨다. 딥러닝은 기계학습 모델에서 인공적인 신경 네트워크 기반의 표상학습으로 볼 수 있으며, 정보처리와 생명공학적인 체계에서 영감을 받아 설계된 개념이다. 인공 신경망 형태로 여러 레이어에 걸친 구조적 디자인을 가진 형태로 대규모 training data를 학습하여 처리함에 있어서, 전통적으로 Empirical Risk Minimization (ERM)이란 형태의 원칙으로 행해져 왔다. ERM은 곧 training data를 이용하여 average error를 최소화시키는 방향으로 전개해 나가며, 충분히 큰 규모의 학습 모델에 대해 어느 정도의 memorization을 허용하고 모델에 대한 generalization을 높이려는 노력을 해왔다. 즉, 이런 ERM의 형태는 다시 말해 한정적인 training data 를 활용하는 것 안에서 input feature vector X와 target vector Y에 대해 예측값과 레이블에 대해 얼마나 차이를 가지는 지 loss function으로 계산하는 empirical risk를 줄이도록 학습하는 데에 초점을 맞춘다. 이에 있어서 발생하는 세 가지 문제가 바로 Memorization, Adversarial Attack, 그리고 Bias이다. 앞선 두 가지의 문제는 머신러닝에 있어서 중요한 테크닉 중 하나인 Mixup으로 알려진 데이터 증강 기법으로 해결

가능하다. Mixup 기법이란 input data x와 class label y 각각에 대해, 무작위로 추출된 두 data에 대해 람다를 계수로 convex combination을 취하는 것에 기반한다. 이러한 과정을 통해 알파 블렌딩 효과를 얻어 training sample에 대한 주변부 학습을 하듯이 다양한 복잡도를 가진 주변부의 무수히 많은 수의 이미지를 활용할 수 있게 되어 데이터 증강 효과를 얻는다. 첫 번째 과제였던 Memorization의 경우, training data에 대해 generalization이 떨어지고 memorize하게 되어 training data 이외에 대해서 성능이 떨어지는 overfitting이 일어나는 것이 문제였는데, mixup을 통한 데이터 증강으로 생성한 거의 무한한 data를 통해 ERM에 비해서 유의미한 training error에 대한 감소를 보인다. 그리고 ERM은 정답 이미지에 대해 특정 악의적인 노이즈를 결합함으로 잘못 판단하도록 유도하는 Adversarial Attack에 취약하였는데, mixup 기법을 사용하면 discriminator에 대한 gradient regularizer 역할을 수행하면서 결과에 대해 loss gradient의 크기를 줄여줌으로 robustness를 향상시키는 효과를 갖는다. Robustness가 향상된다는 것은 잘못된 클래스 레이블로 오인할 가능성을 줄여주는 형태로 defense를 할 수 있다.

오늘 이 강의에선 특히나 세 번째로 언급한 Bias에 대해서 더 중점적으로 설명하셨는데, 학습을 하는데에 중점을 둘 feature가 아닌 엉뚱한 것에 대해 fitting이 이루어지면서 문제가 발생하는 것을 막기 위해 debiasing이 필요한 것이다. 강의에선 Contradicting-pair sampling을 통한 SelecMix 기법에 대한 중점적을 설명하셨다. 간단히 요약하자면, 두 가지 접근법으로 클래스가 동일하고 bias가 다른 것을 각각 선택하여 bias-conflicting data에 대해 data augmentation을 진행하고, 또 bias가 같고 클래스가 다른 bias-conflicting data에 대해 data augmentation을 진행하는 방식이었다. Bias의 유사성에 대한 측정을 통해 auxiliary 모델을 만드는 데에 기반한다고 볼 수 있다.

이번 강연은 확실히 학부에서 알 수 있는 내용 이상으로 새롭게 개념을 알 수 있는 기회를 주었다. 여러 개념에 대해 생소하여 난이도는 있었지만 강연 중 박사님께서 말씀하신 한 가지 내용이 인상적이었다. 무릇 사람이 의사결정하는 데에 있어서 기억하기 쉬우면서 인지적인 부하가 적은 추론에 용이한 방향으로 진행하는데, Bias의 속성이 마치 이에 부합하듯이 신기하게도 학습하는 데에 쉬운 특징을 갖고 있다는 점이었다. 그리고 의도적으로 bias를 학습하는 것을 통해 선별적인 mixup기법 적용으로 해결한다는 점이 흥미로웠다. 더불어서 궁금해지는 점은 바로 필자가 이전에 딥러닝과 관련한 TED 강의를 시청한 것과 관련한데, 해당 내용은 물고기에 대한 image classification을 진행했는데,

자꾸만 사람의 손을 물고기로 인식하는 오류가 발생했다고 한다. 이 원인을 살펴보니, training data에 있던 학습 이미지 중 상당수가 낚시꾼들이 큰 물고기를 잡고나서 자랑하고자 찍은 사진이 포함되어 있어, 학습 과정에서 사람의 손이 물고기 신체의 일부로 인식했다는 류의 내용이었다. 이 경우 Bias 역할을 사람 손이 했다고 볼 수 있는데, background bias와 달리 이 경우도 사람 손에 대한 의도적인 학습을 통해 debiasing을 하는 지 궁금해진다. 이전에는 image processing의 masking과 결부시켜서 생각을 해보았는데, SelecMix 기법을 알게 된 지금이라면, 온전한 물고기 이미지 데이터와 사람 손이 존재하는 데이터에 대해 선별적인 Mixup으로 해결하면 되지 않을까 싶다.