컴퓨터학콜로키움 **(COSE-405) :: 9**주차 노병석 교수님

카카오브레인에서 개발 중인 이미지 및 텍스트 기반 딥러닝 기술

강의 : 2022-11-02 / 작성 날짜 : 2022-11-02

고려대학교 컴퓨터학과 2017320108

고재영

인공지능이란 개념이 우리네 삶에서 긴밀하게 연결된 것은 최근 들어서 비단 놀라운 일은 아니다. 아주 오래 전 원초적이고 상당히 저수준의 반복적인 단순노동을 대신 해주는 기계의 시작에서부터 점점 고차원적이고 높은 수준의 영역에까지 그 확장을 넓혀왔다. 특히나 다중 레이어로 인간의 뇌 신경세포 구조로부터 착안하여 모방한 인공신경망 구조의 딥러닝에 주목해 볼 필요가 있다. 2016년에 비단 대한민국 국내에서 뿐만 아니라 전세계적으로 큰 이슈였던 바둑기사 이세돌과 구글 딥마인드의 알파고와의 바둑 대국 대결이 대중들에게 가장 유명한 딥러닝 시프트의 사건으로 볼 수 있다. 6년 가량의 시간이 흐른 지금은 이보다도 훨씬 더 발전해 왔다. 오늘 이 시간에는 카카오브레인에서 근무 중이신 노병석 연사님께서 카카오브레인에서 개발 중인 이미지 및 텍스트 기반 딥러닝 기술에 대하여 강연을 하셨다.

강연 시작에 앞서서 연사분께서는 하나의 시를 보여주셨다. 그 시는 다름이 아니라 바로 카카오브레인의 한국어 특화 AI 언어모델인 KoGPT를 통해 작성된 시이다. 이 KoGPT는 OpenAI의 딥러닝 AI 모델 중 자연어 기반 소통에 특화된 언어모델 GPT-3을 기반으로 카카오브레인에서 개발한 한국어 특화 모델이다. 흥미로워서 필자가 조금 더 찾아본 결과, 해당 모델은 굉장히 인상적인 점을 가지고 있었다. 기존 딥러닝 방식을 생각해본다면, 감정 분석이라는 태스크를 생각해 볼 때 감정이라는 가치를 사람의 선호도로 대응하여 이를 파악하고자 한다면 수많은 데이터를 확보한 후 데이터들에 대해 좋고 나쁨의 가치 판단을 레이블링을 대응시켜 줘야하는 매우 소모적인 학습 과정이 요구되었다. 하지만 이 GPT 모델은 위에서 언급한 소모적인 추가 학습과정이 필요없이 스스로 판단하여 감정 분석 과정을 해결할 수 있다는 점이 고무적이었다. GPT 모델은 텍스트를 보고 해당 글 내용을 통해 호불호와 같은 긍정/부정을

판별할 뿐만 아니라 내용에 대한 짧은 요약, 그리고 심지어는 인과관계에 대한 예측을 통해서 다음 이야기를 스스로 생성해 낼 수 있다고 한다.

노병석 연사분께서 말씀해주신 강연에서 또 인상적인 부분은 CLIP에 관한 것이었다. CLIP은 Contrastive Language-Image Pretraining의 약자로 최근 OpenAI에서 발표한 모델이다. 기존의 컴퓨터 비전 영역에 있어서 이미지 사전 학습에서 주로 ImageNet의 데이터셋을 주로 사용했는데, 주어진 이미지에 대해 해당 이미지가 보여주는 오브젝트의 종류를 맞추는 일이었다. 하지만 이러한 영상 분류에서는 물체에 대한 레이블링을 직접 연결시키는 소모적인 학습과정이 요구되었고, 이런 이미지로부터 뽑아낸 레이블은 단순히 오브젝트의 종류란 단일 정보만 표현하기 때문에 다양한 특징 중 하나의 feature만 활용하는 한계점을 가진다. 하지만 이 CLIP이란 개념은 데이터의 차원을 줄이면서도 유의미한 형태로 변환해주는, raw text로부터 사전학습을 하는 representational learning에 기반한다. CLIP은 텍스트에 포함된 supervision을 학습하기 때문에, 앞서 말한 기존 방식의 한계점인 레이블링 과정을 생략할 수 있기 때문에 매우 용이하다. 뿐만 아니라 언어에 대한 표현까지 학습을 하기 때문에 유연한 zero-shooting transfer를 가능하게 하는 장점도 존재한다.

마지막으로 소개해주신 내용은 Open-Vocabulary Object Detection에 관한 것이었다. 기존에 오브젝트 탐지하는 방식은 classification과 localization을 기반으로 이미지 영역을 분할하여 적용하는 방식이었다. 해당 방식이 가지는 가장 치명적인 단점은 기존에 predefined되어 있는 오브젝트에 대해서만 검출이 가능한 것이었다. 따라서 미리 정의되어 있는 오브젝트 레이블링 뿐만 아니라 범물체적으로 잘 검출하기 위해 제안된 Open Vocabulary Object Detection 모델은 앞서 언급되었던 CLIP이란 패러다임이 적용되었다. CLIP 모델을 통해 사전 학습을 거치는 text encoder 과정을 거쳐 text embedding에서 결과를 뽑아내어 굉장히 효과적으로 꾀했다고 한다. 가장 대표적으로 적용할 수 있는 실생활 분야는 감시 시스템이나 이미지 검색으로 볼 수 있다. 사람의 소모적인 노동을 절감하며 구체화된 쿼리를 통해 자동 분석이 가능하기 때문에 굉장히 혁신적인 결과를 가져올 수 있는 것이다. 개인적으로 이미지 검색이라는 꽤 범용적인 면에 적용할 수 있는 부분이 고무적으로 다가왔다. 필자도 온라인 쇼핑을 자주 하는 편인데 기존에 검색 패러다임을 바꾸어 훨씬 효과적이고 간편하게 쿼리를 통해 사용자가 원하는, 어쩌면 사용자조차도 명확히 몰랐던 타겟을 검색 가능하게 할 수 있을 것 같다.