2017320114 최재원

고수준 장면 이해를 위한 트랜스포머 기반 사람 사물 상호작용 탐지 및 MLV Lab 최신 연구

2017320114 컴퓨터학과 최재원

현재에서는 Object detection, 즉 그 장면 속의 물체를 인식하는 것은 아주 잘되고 있지만, 사실은 어떤 장면에 어떤 물체가 있는 것만을 가지고는 그 장면을 깊이 있게 이해한다고 볼 수는 없다. 그래서 그 장면 속에 물체들이 어떠한 상호작용을 하고 어떤 의미를 갖는지에 대한 사람과 사물과의 상호작용을 찾는 알고리즘에 대해서 김현우 교수님의 MLV Lab 최신 연구와 함께 배울 수 있는 시간을 가졌다. 또한 나로서도 Object Detection에 대해서는 많이 들어봤지만 거기서 더 나아가서 고수준 장면 이해라는 주제를 들었을 때 매우 흥미롭게 다가왔다.

고차원 장면 이해를 하는 기본적인 방법은 이미지가 있으면 Object detection을 먼저 진행을 한다. 그리고 이 Object들간의 relationship이 무엇인지 모든 Pair로 예측을 진행한다. 예측은 Graph Network 등 다양한 방법으로 진행을 한다. 모두 진행하고 나면 한 이미지에서 모든 detection된 bounding box가 나오고 Scene Graph라는 것을 만들게 된다. 하지만 이러한 것들을 다 detection을 하고, 딥러닝 feature들을 뽑아내고 하는 예측 과정이 너무 오래 걸려서 UnionDet이라는 모델을 만들게 되었다. 이것은 detection을 한 다음에 관계를 찾는게 아니라, detection은 detection대로, 관계는 관계대로 병렬적으로 연산을 한 다음에 두 개를 묶어서 대답을 하면 엄청 빨리 된다는 것이 아이디어이다. UnionDet을 이용해서 다양한 거리의 상호작용 탐지도 가능하고 feature들끼리 뒤섞여서 오류가 나는 것도 줄어들었다.

HOTR(End to End Human-Object Interaction Detection with Transformers)란 트랜스포머라는 딥러닝, 언어, 비전 등에서 유명한 모델을 이용해서 Human Object Interaction Detection을 푸는 문제이다. 이 문제는 Object Detection과는 다르게 여러가지 레이블을 한꺼번에 예측을 해야 된다는 어려움이 있다. 기본적으로는 Detection을 한 다음에 두 개 마다 어떤 관계인지 feature를 다시 뽑아서 순차적으로 하기 때문에 시간이 오래 걸린다. 따라서 이렇게 하는 법 대신에 트랜스포머를 이용해서 병렬적으로 한꺼번에 할 수 있게 의도를 하는 방법이다. 먼저 이미지가 있으면 Convolutional Neural Network을 이용해서 feature를 뽑아내고 각 feature들이 어떠한 위치에서 왔는지 알기 위해

서 트랜스포머인 기본적인 데이터 처리 방법인 Positional Encoding을 만들어서 넣어준다. 그 다음에 이걸 쪼개서 transformer encoder에 넣어서 encoding을 해준다. 이 정보를 다시 object detection을 위해서 Instance Decoder에다가 우리가 쿼리를 가지고 있고 또 interaction decoder HOI union을 통째로 detection 하던 UnionDet의 아이디어를 이용해서 detection을 해준다. Decoding을 해주게 되면 이 정보들이 FFN을 거쳐서 어떤 detection 결과인지 class, bounding box를 통해서 나오고 interaction decoder를 통해서는 어떤 물체들이 interaction을 하고 있는지 예측하는 결과가 나온다. 결론적으로는 Decoding 된 결과가 다른 decoding 된 pointing하는 포인터들을 예측하고 그리고 그 두 개의 관계를 Multi label classification을 분류하는 문제로 예측을 해서 푸는 것이다. 여기서 흥미로운 것은 HO Pointer인데, HO pointer를 이용해서 일반적인 Object detection과 Human Object interaction detection과 연결고리를 만들어준 것이다. 효율적으로 병렬적으로 하면서도 두 개가 같이 움직일 수 있도록 Pointer로 만들어줘서 성능을 상당히 올린 것이다.

현재 컴퓨터 비전은 2D가 아니라 3D 나아가서 4D까지 되어있다. Point Cloud를 이용한 이미지인데 시간에 따라 변화를 하는 것을 어떻게 잘 처리하고 이해를 할 지에 대해서 연구는 진행중이다. 딥러닝 모델들이 점점 커지지만 데이터의 양은 아무리 많아도 한정적이다. 어떻게 하면 효과적으로 딥러닝 모델을 3D 데이터에서 잘 학습시킬 수 있을까라고 했을 때 가장 많이 쓰는 방법 중에 하나인 데이터 증강 기법을 개발을 했다. 사실 데이터 증강 기법이 point cloud를 위해서는 많이 없었다. 이것에 대한 교수님이 만든 연구의 기법은 Point Wolf라고 해서 Weighted Local Transformation을 줄인 것이다. Point Wolf란 각 위치마다 조금씩 돌리거나 찌그러트리거나 늘려서 가만히 있는 wolf를 뒷다리를 빼게 하거나 점프를 하게 한다던가 있을 법한 샘플들을 만들었다. 따라서 이러한 동작을 해도 늑대라는 것을 인식시켜준다.

이렇듯 이번 세미나를 통해서 내가 알고 있는 Object Detection을 넘어서 Detection 된 object 및 human과의 관계까지 파악하는 법에 대해서 배울 수 있게 되었다. 흥미로웠던 점은 교수님의 연구실에서 개발한 기법들이 현재 제일 효율적으로 나와있다는 점과 현재 진행형이라는 것이었다. Machine Learning에 관심이 많은 학생으로서 물체들 간의 관계를 파악하고 장면을 이해한다라는 것이 매우 신기했고 스스로 좀 더 알아보고 싶을 만큼 유익한 시간이었다. 기회가 된다면 앞으로 이러한 연구의 일원이 되어서 AI의 발전에 이바지를 하고 싶다.