언어지능의 현황, 한계, 도전

2017320114 컴퓨터학과 최재원

컴퓨터 비전과는 달리 언어 쪽에서는 인간을 뛰어넘기가 어려울 것이라고 많은 사람들이 생각을 했다. 기술이 실제로도 안 나오기도 했다. 그러다가 2018년부터 세상이 바뀌면서 지금은 비단 언어만이 아니라 비전, 음성 쪽에서도 BERT의 basic building block 인 transformer를 가져 다가 응용을 하고 있다.

그렇다면 어떻게 컴퓨터가 워드(word)를 표현할까? 일반적으로 가장 최근까지 사용한 기법은 one hot encoding 기법이다. 이 기법은 각자의 identity를 가지고 있는, 예를 들어 단어가 만약에 백만개가 있고 apple을 찾고 싶으면 백만 개 중에 하나만 1인 그러한 기법이다. 우리가 알고 있는 아스키 코드 같은 것들이 대표적인 one hot encoding 기법이다. 그러다가 늦게 꽃 핀 기법이 distributed representation이다. 이 기법은 단어들이 연속되는 activation level에 따라 표현되는 것이다. 딥러닝 등에서 사용되고 있다. Distributed representation이 one hot encoding이랑 다른 점은 one hot encoding에서는 전체에 있는 단어들 중에서 apple에 해당하는 위치만 식별하는 것이다. 즉 이 자리에 1이 있으면 이것은 apple이야라는 것이다. 그에 반해서 neural model이 지향하는 세상은 좀 많이 다르다. 예를 들어 apple를 표현하는 neuron이 다섯 개가 있다고 하면, 이것은 마치 인간이 심상을 떠올리는 것처럼 apple이라고 하면 뭔가 neuron의 집합으로 표현이 되고 orange도 비슷한 패턴을 보이는, 반대로 car라고 하면 다른 패턴으로 보이는 이러한 인간이 인지하는 방법이랑 비슷하게 가는 방법이다. 이러한 방법들 중에 유명한 analogy인 word2vec은 워드를 벡터로 바꾸는 것이다. Word2vec은 주위의 있는 단어들 하고 뜻이 비슷하다고 가정한다. 이것을 우리가 distributional hypothesis라고 하는데 이것을 구현하는 모델을 Word2vec이라고 하고 엄청난 성공을 거두었다. 이것을 가능케 했

던 것은 엄청난 양의 데이터를 학습시켰기 때문이다. 그 이외에도 Doc2vec 등이 있다.

지금까지 봤던 것들은 shallow model이다. 즉 hidden layer가 하나밖에 없다는 뜻이다. Input과 output을 제외하면 hidden layer가 하나밖에 없다. 그 뒤로는 이제 deep neural network이 들어오게 되고, CNN을 이용한 언어처리도 가능하다는 것을 알아내게 되었다. Transformer가 오기 전에는 언어처리를 CNN이나 RNN을 사용해서 했다. 하지만 Transformer의 등장 이후 언어 처리의 형태가 사뭇 달라졌다. Transformer란 구글에서 2017년에 발표한 것이다. Transformer의 구조를 뜯어보면 attention밖에 없다는 것을 알 수 있다. Transformer는 처음에 machine translation을 위해서 만들어졌다. Machine translation을 하려면 하나의 언어를 다른 언어로 바꾸어야 되니까 input으로 들어온 언어를 encoder로 해석하고 출력으로 나오는 언어를 decoder로 해석하면 된다. 다만 Decoder 쪽에서는 attention block이 하나가 더 있다. 여기서 Bert는 transformer 모델에서 encoder만 따와서 쌓아 올린 모델이고, G PT(1,2,3)는 transformer에서 decoder만 가져와서 만든 모델이다. 마지막으로 최근에 인기 많은 바트라는 모델은 인코더와 디코더를 둘 다 쓴다. 이게 지금 오늘날의 AI의 핵심 기술이다.

이상근 교수님의 언어지능의 발전 및 기술 설명 수업을 듣기 이전에는 사실 언어 처리에 대해 큰 관심을 가지고 있지는 않았다. 하지만 이번 강의를 통해서 최신 경향 및 앞으로의 발전 가능성에 대해서 알 수 있게 되었고, 이전보다 많은 흥미를 가질 수 있는 수업이 되었던 것 같다. 앞으로 컴퓨터 비전과 함께 같이 성장할 수 있었으면 좋겠다.