Cs231n Lecture 5 Summary
Convolutional Neural Networks

1. **History & The use of CNN**
2. **Convolutional Neural Networks**

1. History & The use of CNN

1957 년 최초로 perception 이 구현되었다. (Mark I perception) 여기서 가중치 W 를 업데이트하는 update rule 이 나타났다. 그리고 1960 년에는 Multilayer Perception Network 인 Adaline/Madaline 이 발명되었다. 1986 년 처음으로 backpropagation 이 나타났고 신경망의 학습이 시작되었다. 그러나 여러 가지 상황으로 인해 암흑기에 빠지게 되었다.
다시 2006 년, Deep Learning 의 학습 가능성이 보이기 시작했다.
초기에는 RBM 이 이용되었고 hidden layer 에서 가중치가 학습되었다. 그 후 전체 신경망을 backpropagation 하거나 fine-tuning 하는 방식으로 진행되었다. 신경망의 광풍은 2012 년부터 본격적으로 시작되었다. ImageNet 분류를 통해 오차를 극적으로 감소시키는 AlexNet 이 나타났고 이때부터 CNN 이 본격적으로 사용되기 시작했다.
CNN 의 역사는 1959 년부터 시작한다. 처음에는 고양이의 뇌에 신호로 자극을 주었을 때 고양이가 edges 와 shapes 에 반응을 보인다는 사실을 알게 되었다. 그 이후 뇌의 특정 뉴런은 특정 방향에 반응한다는 것을 알게 되었다. 그리고 뉴런이 계층 구조를 가진다는 것을 알아냈는데, Retinal ganglion cell 을 통해 시각 정보를 받아들일 때 반응을 보이는 simple cells 로부터 complex cells, corner, blob 등 hypercomplex cells 의 계층 구조를 가진다.
1980 년 Neocognitron 이 탄생하였는데, 이러한 생물학적 성과들을 컴퓨터로 시뮬레이션 한 결과이다. Simple cell 과 complex cell 을 반복적으로 쌓는 샌드위치 구조를 가지고 있다. 하지만 backpropagation 은 하지 못했다.
1988 년 gradient-based learning 이 글자 인식에 적용되었다. 이것은 우편번호의 숫자를 인식하는 용도로 사용되었다. 이 모델은 backpropagation 이 가능하다.
마침내 2012 년, AlexNet 이 나타났다. 1988 년에 나온 것과 구조는 같지만 더 크고 깊어진 것이 특징이다. 대규모의 데이터를 활용하였고 처음으로 GPU 를 2 대 사용하였다. 또한 가중치 초기화, batch normalization 을 하였다.

CNN 은 다양한 분야에 사용된다. 분야에는 이미지 분류, 이미지 검색, Detection, segmentation, 자율 주행, 비디오 인식, face recognition, pose recognition, 의학 영상 해석, 게임 등이 있다.

2. Convolutional Neural Networks

이전 강의에서는 FC Layer에 대해 배웠다. FC Layer는 32x32x3의 이미지를 늘려서 3072x1의 벡터 x로 만들어 가중치 W와 내적을 하여 1x10의 activation layer에 출력한다.

그러나 Convolution Layer는 기존 이미지의 구조를 유지하고 filter와의 공간적 내적을 통해 1개의 숫자를 출력한다. Convolution Layer에서는 filter의 크기를 정할 수 있는데, 가로와 세로는 선택 가능하지만 (ex. 5x5) depth는 입력 이미지의 depth와 같아야 하므로 filter의 크기는 5x5x3이 된다. 이처럼 입력 이미지에 filter를 슬라이딩하여 내적을 구하는 것을 convolve한다고 한다. 이렇게 convolve를 하면 1개의 결과 숫자가 나온다. 이 숫자는 convolve한 개수만큼 나오는데, 만약 10번 convolve를 한다면 10개의 숫자가 나온다. 이 숫자가 나오는 식은 w^T*x + b이다. 여기서 w^T*x는 5 * 5 * 3 = 75이고, b는 bias이다. 이 5x5x3의 filter가 32x32x3의 이미지를 1번 슬라이딩하면서 내적해 얻은 값을 모두 모으면 28x28x1의 이미지를 얻게 되는데, 이를 activation map이라고 한다. 출력된 activation map의 크기는 filter의 크기와 숫자, 슬라이딩 방식에 따라 달라진다.

각 필터는 이미지로부터 1개의 특징을 추출하는데, CNN에서는 여러 개의 필터를 사용하므로 이미지로부터 여러 개의 특징을 추출해 낼 수 있다.

CNN에서는 입력 이미지가 convolution layer와 활성함수 ReLU를 통과하여 activation map을 얻고 이 activation map에 다시 convolution layer와 활성함수를 통과하는 과정을 반복적으로 하게 된다. CNN 모델 디자인은 필터의 크기와 숫자, stride를 통해 다양하게 선택될 수 있다.

필터가 여러 개일 때, 필터는 이미지의 특징을 단순한 것부터 복잡한 것까지 계층적으로 학습하게 된다. 이는 CNN이 1959년 발견한 계층 구조를 가지는 뉴런과 유사하다는 것을 보여준다.

슬라이딩을 할 때, 불균형한 정보를 가진 activation map이 나올 수 있다. 이때는 zero padding이라는 방법을 사용하는데, 이미지의 바깥에 0으로 이루어진 픽셀을 붙여서 해결한다. 이러면 이미지의 가장자리 정보도 출력 이미지에 잘 전달할 수 있게 된다.

32x32의 이미지에 5x5의 필터로 반복하여 슬라이딩하면 activation map의 크기는

빠르게 줄어들 것이다. 이는 빠르게 정보가 손실됨을 의미하므로 정확한 정보를 추출할 수 없다. 그래서 convolution layer에서는 zero padding을 하여 출력 이미지의 크기를 보존하고 이미지의 크기를 줄이는 작업은 따로 pooling에서 진행하게 된다.

Convolution layer는 입력 이미지를 국소적으로 여러 번 보고, FC layer는 이미지를 전체적으로 한 번 보는 것과 같다. 따라서 convolution layer에서는 이미지를 확대, 축소, 이동하더라도 이미지의 특징을 잘 찾을 수 있지만 FC layer에서는 이미지 전체의 특징 하나를 추출하므로 효과적이지 않다.

Pooling layer는 표현을 작게 만들어(downsampling) 관리하기 쉽게 만들고 각 activation map에 독립적으로 작용한다. 즉 이미지의 차원을 공간적으로 줄여 준다. 하지만 depth는 줄이지 못한다. 가장 많이 사용되는 pooling 방법은 max pooling인데, 필터의 크기와 보폭을 이용하여 downsampling을 한다.