Cs231n Lecture 7 Summary
Training Neural Networks 2

1. **Optimization**
2. **Regularization**
3. **Transfer learning**

1. Optimization

Neural network 에서 가장 중요한 것은 최적화 문제이다. 지금까지 배운 최적화 알고리즘 중 가장 간단한 알고리즘은 gradient descent(경사하강법)이었다. 이는 batch size 로 나눠둔 모델에 가중치를 곱해서 loss 를 줄이는 방법이다.

1.1. SGD

미니 배치 안에서 loss 를 계산하는 방법이 SGD 이다. SGD 는 gradient 의 반대 방향을 이용해 파라미터 벡터를 업데이트한다. 하지만 이 방법에는 문제점이 있는데, loss 는 수직 방향의 가중치 변화에 훨씬 더 민감하게 반응한다. 수평일 경우에는 아주 미세하게 줄어든다. 따라서 loss 는 bad condition number 를 가진다고 말할 수 있다.

Loss 가 원점으로 가기. 위해서 수직적 이동을 해야 하고, 따라서 zig-zag 모양으로 이동해 효율이 떨어진다.

SGD 의 문제점은 크게 두 가지가 있는데, local minima 문제와 saddle point 문제이다. 둘 다 SGD 의 특성으로 인해 gradient 가 0 이 된다.

이를 보완하기 위해 나온 방법이 SGD+Momentum 이다. 식에 Vx(velocity)가 추가되어 local minima 에 도달하여 gradient 가 0 이 되어도 계속 움직인다. 따라서 momentum 을 추가하면 loss 에 민감한 수직 방향의 변동을 줄일 수 있어 좋다.

1.2. AdaGrad, RMSProp, Adam

AdaGrad 는 훈련 도중 계산되는 gradients 를 활용하는 방법이다. Velocity term 대신에 grad squared term 을 이용하고 update 를 할 때 update term 을 앞서서 계산한 gradient 제곱으로 나눠준다. Gradient 가 작은 값일수록 제곱시에는 더 작아지기 때문에 가속도가 붙는다. 단점은 진행할수록 값이 점점 더 작아진다. 따라서 step size 를 더 작게 해주어야 한다. 이는 non-convex cads 에서는 멈춰버리기 때문에 나쁜 특징이다.

RMSProp 는 AdaGrad 의 변형 모델이다. AdaGrad 의 gradient 제곱을 그대로 사용하지만, 기존의 누적값에 decay_rate 를 곱해준다. 따라서 gradient 의 제곱을 계속해서 누적한다. Decate rate 값은 0.9 나 0.99 가 주로 쓰인다.

Adam 은 momentum 과 RMS 를 합친 모형으로 더욱 개선된 모형이다. 1 차, 2 차 momentum 으로 작동한다. Global minima 를 찾다가 최솟값을

뛰어넘을 수도 있기 때문에 신중하게 작동한다. 빠르게 수렴한다는 장점이 있지만 계산 비용이 많이 든다는 단점이 있다.

1.3. Model Ensembles

모델 앙상블은 모델을 하나만 학습시키는 것이 아닌 여러 가지 다른 모델을 결합하는 머신러닝 접근 방식이다. 단일 모델로는 높은 정확도를 가지기 어렵지만, 여러 모델을 만들고 결합하면 전체 정확도가 향상될 수 있다.

2. Regularization

2.1. Dropout

Dropout 은 Neural Network 에서 가장 많이 사용하는 Regularization 방식이다. Forward pass 과정에서 임의로 뉴런을 0 으로 만든다. Forward pass 를 진행할 때마다 0 이 되는 뉴런이 바뀐다. Dropout 은 한 Layer 씩 진행한다.

Dropout 은 일부 값들을 0 으로 만들면서 네트워크를 훼손시키는 것처럼 보이지만 그렇지 않다. 특징들 간의 상호작용을 방지하여 overfitting 을 막는 것에 효과가 있다. 그래서 dropout 은 단일 모델로 앙상블 효과를 가질 수 있다.

2.2. Data augmentation

또 다른 방식은 data augmentation 이다. 원래의 학습 과정에서는 데이터와 레이블이 있고 이를 통해 매 스탭마다 CNN 을 업데이트하였다. 하지만 대신 train time 에 이미지를 무작위로 변환시켜 볼 수 있다. 그렇다면 원본 이미지의 label 을 훼손시키지 않으면서 다양한 경우의 학습을 진행 가능하다.

3. Transfer learning

CNN 에는 엄청나게 많은 데이터가 필요하다고 생각할 수 있지만, transfer learning 은 그러한 생각을 깨뜨린다. 보통은 overfitting 의 원인을 소규모

데이터라고 말하였는데, 소규모 데이터를 학습하면서도 transfer learning 방식을 이용해 overfitting 을 방지할 수 있다.
먼저 CNN 을 가지고 아주 큰 데이터셋(ImageNet)으로 학습을 시킨다. 그리고 여기서 학습된 feature 를 우리가 가진 소규모 데이터셋에 적용시킨다. 여기서 가장 마지막의 FC layer 를 초기화시키고 오로지 마지막 레이어만 가지고 데이터를 학습시킨다.