6. 배치 정규화

Normalizing Activations in a Network

Batch Normalization

- 하이퍼파라미터 탐색 쉽게 함
- 깊은 심층신경망도 쉽게 학습할 수 있게 함

Batch Normalization in Hidden Units

- activation function 이전에 사용
    - $\mu = \frac{1}{m} \sum_i z^{(i)}$
    - $\sigma^2 = \frac{1}{m} \sum_i (z^{(i)} - \mu)^2$
    - $z_{norm}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$
    - $\tilde{z}^{(i)} = \gamma z_{norm}^{(i)} + \beta$

- hidden units이 평균0 표준편차1을 갖는 것은 좋지 않기 때문에 $\tilde{z}^{(i)}$을 사용함
- γ, β는 학습해아하는 파라미터
- 입력층과 은닉 유닛의 차이점
    - 평균과 분산이 각각 0, 1이 아닌 다른 값을 가짐
    - 표준화된 평균과 분산을 갖되 γ, β에 의해 조절됨

Fitting Batch Norm Into NN

1st layer : $X \xrightarrow{w^{[1]},\, b^{[1]}} z^{[1]} \xrightarrow[\text{Batch Norm}]{\beta^{[1]},\, \gamma^{[1]}} \tilde{z}^{[1]} \longrightarrow a^{[1]} = g^{[1]}(\tilde{z}^{[1]})$

$\downarrow\ W^{[2]},\, b^{[2]}$

2nd layer : $z^{[2]} \xrightarrow[BN]{\beta^{[2]},\, \gamma^{[2]}} \tilde{z}^{[2]} \longrightarrow a^{[2]}$

1) Z 계산한 후 이를 Batch Norm을 통해 $\tilde{z}$ 구함

2) $\tilde{z}$ 값을 activation function을 거쳐 a를 구함

Implementing gradient descent

- for t=1… #mini-batches
    - Compute forward prop on X[t]
        - In each hidden layer, use BN to replace Z[l] with $\tilde{z}^{[l]}$
    - Use backprop to computes $\delta w^{[l]}, \delta\beta^{[l]}, \delta\gamma^{[l]}$
    - Update parameters $w^{[l]}, \beta^{[l]}, \gamma^{[l]}$

Why Does Batch Norm Work?

- 신경망에서 더 깊은 층으로 갈수록 가중치 변화에 영향을 덜 받음
- Coveriate Shift의 영향을 줄여줌
    - Coveriate Shift
        - 이전 layer의 파라미터 w, b가 바뀌면 그에 따라 다음 layer의 input이 되는 a가 바뀌면서 데이터의 분포가 바뀌게 됨
        - X의 분포가 바뀌면 모델을 다시 훈련시켜야함
- Regularization effect
    - 미니배치로 계산한 평균과 분산은 noise가 있음
    - 곱셈 잡음과 덧셈 잡음이 모두 있어 regularization 효과가 있음
    - 미니배치 크기가 커질수록 정규화 효과는 감소함

Batch Norm at Test Time

- test 는 배치가 하나임 → 평균과 분산을 계산할 수 없음
- 학습시에 사용된 미니배치의 지수가중평균을 추정치로 사용