

6. 배치 정규화

날짜 @2023년 11월 25일

▼ 목차

[배치 정규화](#)

[1 Normalizing inputs to speed up learning](#)

[2 Implementing Batch Norm](#)

[배치 정규화 적용시키기](#)

[1 Adding Batch Norm to a network](#)

[2 Implementing Gradient Descent](#)

[배치 정규화가 잘 작동하는 이유](#)

[1 Why Does Batch Norm Work?](#)

[2 Batch Norm as regularization](#)

[테스트 시의 배치 정규화](#)

[출석퀴즈 오답노트](#)

배치 정규화

1 Normalizing inputs to speed up learning

- 활성화 함수 이전의 입력 변수들을 평균과 분산을 사용하여 정규화함
- 배치 정규화의 장점
 - 하이퍼파라미터 탐색을 쉽게 만들
 - 신경망과 하이퍼파라미터의 상관관계를 감소시킴
 - 더 많은 하이퍼파라미터가 잘 작동하도록 하여 깊은 신경망도 쉽게 학습할 수 있게 도움

2 Implementing Batch Norm

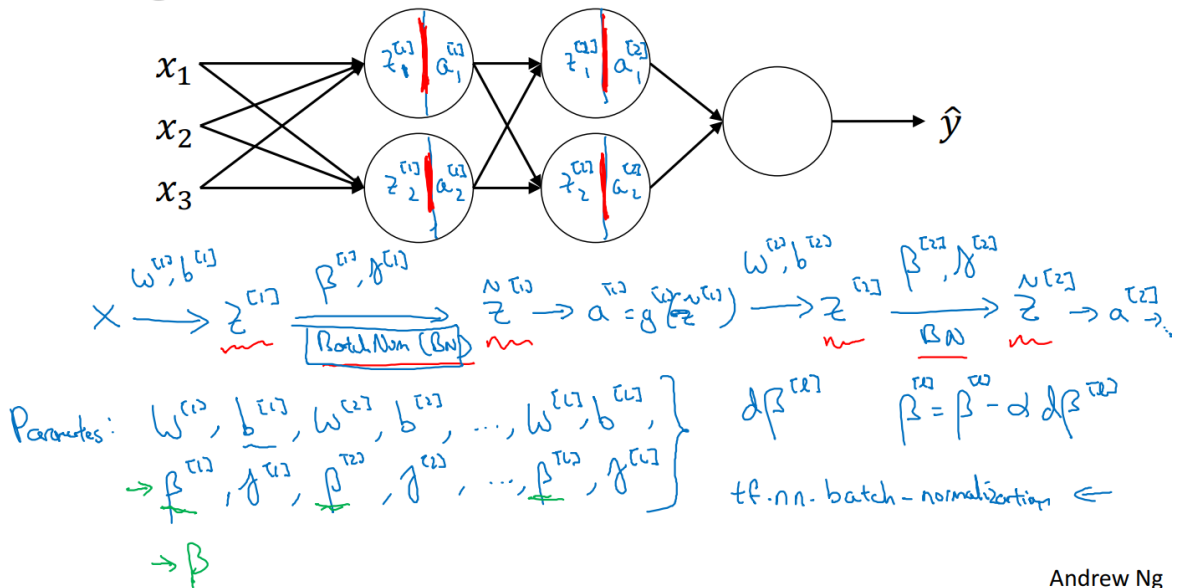
- 배치 정규화의 작동원리
 - $\mu = \frac{1}{m} \sum_i z^{(i)}$ (평균)
 - $\sigma^2 = \frac{1}{m} \sum_i (z^{(i)} - \mu)^2$ (분산)
 - $z_{norm}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$

- ϵ 은 수학적 안정성을 위해 추가됨
- $\tilde{z}^{(i)} = \gamma z_{norm}^{(i)} + \beta$
 - 정규화 이후 다시 선형변환함으로써 은닉 유닛이 같은 분포 값을 갖지 않도록 함
 - γ 와 β 는 학습과정에서 학습하는 파라미터

배치 정규화 적용시키기

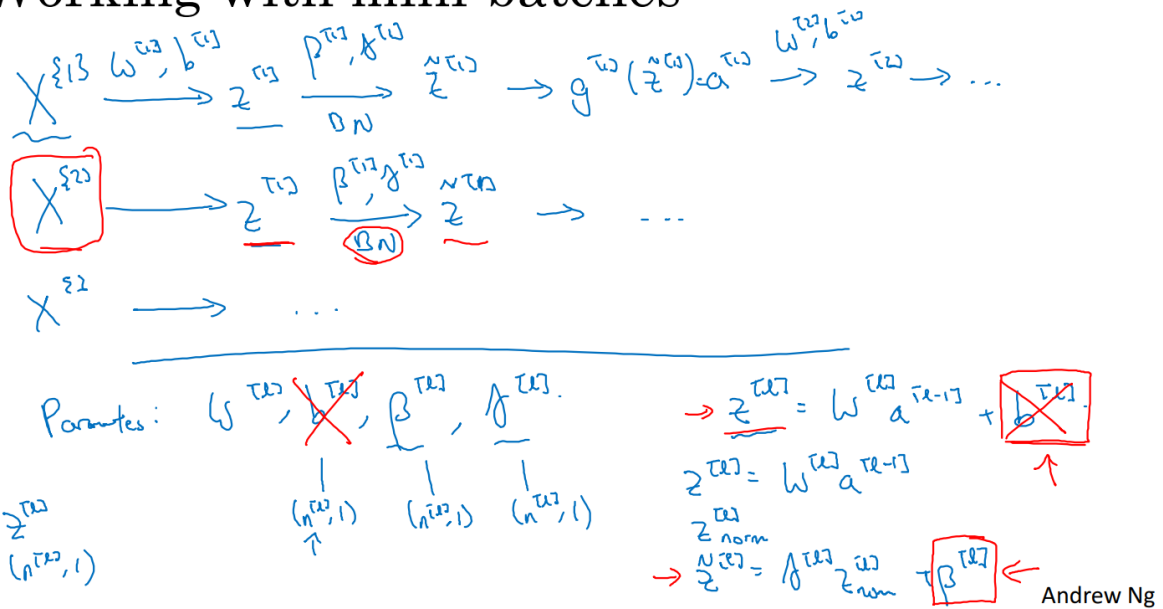
1 Adding Batch Norm to a network

Adding Batch Norm to a network



Andrew Ng

Working with mini-batches



- 다음 과정을 매 은닉층마다 반복함
 - 선형결합인 z 를 계산하고 이를 배치 정규화(BN)시킴
 - β 와 γ 로 \tilde{z} 를 얻음
 - 정규화된 값들을 활성화함수를 거쳐 활성화 값 a 를 얻음
- 선형결합 단계에서 상수항 b 가 없어짐
 - 배치 정규화 과정에서 z 의 평균을 계산한 뒤 빼 줌으로써 사라지기 때문

2 Implementing Gradient Descent

Implementing gradient descent

for $t = 1 \dots \text{num Mini Batches}$
Compute forward pass on X^{set} .
In each hidden layer, use BN to replace z^{set} with \hat{z}^{set} .
Use backprop to compute $\frac{dL}{dW^{\text{set}}}$, $\frac{dL}{d\beta^{\text{set}}}$, $\frac{dL}{d\gamma^{\text{set}}}$
Update parameters $\left. \begin{aligned} W^{\text{set}} &:= W^{\text{set}} - \alpha \frac{dL}{dW^{\text{set}}} \\ \beta^{\text{set}} &:= \beta^{\text{set}} - \alpha \frac{dL}{d\beta^{\text{set}}} \\ \gamma^{\text{set}} &:= \dots \end{aligned} \right\} \leftarrow$
Works w/ momentum, RMSprop, Adam.

Andrew Ng

배치 정규화가 잘 작동하는 이유

1 Why Does Batch Norm Work?

- 배치 정규화는 입력 특성 X 의 평균을 0, 분산을 1로 만듦으로써 학습 속도를 빠르게 함
- 배치 정규화가 잘 되는 이유 중 하나는 이전 층의 가중치 영향을 덜 받게 하는 것에 있음
 - 은닉층 값의 분포 변화를 줄임으로써 입력 값의 분포를 제한하기 때문
 - 즉, 배치 정규화는 입력 값이 바뀌어서 발생하는 문제를 안정화시킴
 - 앞층과 뒷층의 매개변수의 상관관계를 줄여 학습속도를 향상시킴

2 Batch Norm as regularization

- 배치 정규화의 또다른 효과는 파라미터의 정규화임
 - 미니배치로 계산한 평균과 분산은 전체 데이터의 일부로 추정한 것이기 때문에 잡음이 존재함
- 드롭아웃의 경우 은닉 유닛의 확률에 따라 0 또는 1을 곱하므로 곱셈 잡음이 발생
- 배치 정규화의 경우 곱셈잡음($\times \frac{1}{\sigma}$)과 덧셈잡음($+(-\mu)$)이 동시에 발생하므로 약간의 정규화 효과가 존재

- 은닉층에 잡음을 추가하는 것은 이후 은닉층이 하나의 은닉 유닛에 너무 의존하지 않도록 만들
- 큰 미니배치 사용 시 이 정규화 효과는 상대적으로 약해짐

테스트 시의 배치 정규화

Batch Norm at test time

→ $\mu = \frac{1}{m} \sum_i z^{(i)}$

→ $\sigma^2 = \frac{1}{m} \sum_i (z^{(i)} - \mu)^2$

→ $z_{\text{norm}}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$

→ $\tilde{z}^{(i)} = \gamma z_{\text{norm}}^{(i)} + \beta$

μ, σ^2 : estimate using exponentially weighted average (across mini-batches).

$x^{[1]}, x^{[2]}, x^{[3]}, \dots$

↓

$\mu^{[1]}, \mu^{[2]}, \mu^{[3]}, \dots \rightarrow \mu$

$\sigma^{[1]}, \sigma^{[2]}, \sigma^{[3]}, \dots \rightarrow \sigma^2$

$z_{\text{norm}} = \frac{z - \mu}{\sqrt{\sigma^2 + \epsilon}}$

$\tilde{z} = \gamma z_{\text{norm}} + \beta$

Andrew Ng

- 테스트 시에는 배치가 하나이므로 평균과 분산을 계산할 수 없음
- 학습 시 사용된 미니배치들의 지수 가중 이동 평균을 추정치로 사용함

출석퀴즈 오답노트

- ▼ 6. 배치 정규화의 β 와 γ 에 대한 설명으로 옳은 것
- 정규화 이전의 데이터에 사용된다. (X)
 - γ 와 β 를 활용하여 선형변환하면 같은 분포를 갖게 된다. (X)
 - 경사하강법뿐만 아니라 Adam, momentum을 이용한 경사하강법, RMSprop을 사용하여 학습할 수 있다. (O)
 - 주어진 레이어의 선형변수 $z^{[l]}$ 의 평균과 분산을 설정한다. (O)
- ▼ 8. TF

배치 정규화를 사용하여 신경망을 훈련시킨 후 테스트 시 새로운 데이터에서 신경망을 평가하려면, 필요한 정규화를 수행하고, 훈련 중에 확인된 미니 배치에 대한 지수 가중 평균을 사용하여 추정된 μ 및 σ^2 를 사용한다. (O)