Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

Sergey Ioffe, Christian Szegedy

23/03/31

서강대학교 정보통신대학원
배성은

목차

(0) Abstract

✓ Deep Neural Networks 훈련은 이전 레이어의 parameters가 변경됨에 따라 **훈련 중에 각 레이어의 입력 분포가 변경된다는 사실 때문에 복잡하다**.
✓ 이는 낮은 학습 속도와 신중한 parameters 초기화를 요구하기때문에, 훈련 속도를 늦추고, 비선형성 모델을 훈련시키는 것을 힘들게 만듬.
✓ 이 현상을 **internal covariate shift**라고 함.

✓ **레이어 입력을 정규화하여 문제를 해결**한다.
✓ 정규화를 모델 아키텍처의 일부로 만들고 각 훈련 미니 배치에 대해 정규화를 수행함으로써 해결.
✓ 배치 정규화를 사용하면 훨씬 높은 학습 속도를 사용할 수 있으며, 초기화에 주의를 덜 들일 수 있음.
✓ 경우에 따라 Dropout의 필요성을 없애줌.

✓ Batch Normalization은 14배 적은 train으로 동일 정확도를 달성하고, 원래 모델을 능가함.
✓ 앙상블을 사용해선 ImageNet 분류에 대한 좋은 정확도를 냄.

(1) Introduction

- ✓ SGD는 심층 네트워크를 훈련시키는 효과적인 방법으로 입증되었음.
- ✓ SGD는 네트워크의 매개변수를 최적화하여 손실을 최소화함.
- ✓ SGD는 단계적으로 training되며, 각 단계에서 크기m의 미니배치를 고려하는데 미니배치는 매개변수에 대한 loss function의 기울기를 계산하여 사용함.
- ✓ Stochastic gradient는 간단하고 효과적이지만, 모델 매개 변수의 초기값 뿐만 아닌 최적화에 사용되는 학습 속도와 같은 **모델 hyper-parameters의 신중한 튜닝이 필요.**
- ✓ 네트워크 매개변수의 작은 변경은 네트워크가 깊어짐에 따라 증폭.

- ✓ Layer의 입력 분포 변화는 Layer가 새로운 분포에 지속적으로 적응해야하기 때문에 문제가 생김.
- ✓ 학습 시스템에 대한 입력 분포가 변경되는 것을 **covariate shift**라고 함.
- ✓ 일반적으로 도메인 적응을 통해 처리되나 covariate shift 개념은 학습 시스템 전체를 넘어 확장될 수 있음.
- ✓ 하위 네트워크에 대한 입력의 고정 분포는 하위 네트워크 외부 레이어에 긍정적인 영향을 미침.
- ✓ In-ternal Covariate Shift를 제거하는 것은 더 빠른 훈련을 가능하게 함.
- ✓ 매개 변수의 척도 또는 초기 값 그리디언트 의존성을 줄여 흐름에도 유익한 영향을 미침.
- ✓ 모델을 정규화하고 drop out 필요성을 줄임.

(2) Towards Reducing Internal Covariate Shift

- ✓ Internal Covariate Shift as the change in the distribution of network activations due to the change in network parameters during training.
- ✓ (내부 공변량 이동을 학습 중 네트워크 매개변수 변경으로 인한 네트워크 활성화의 분산 변화로 정의)

- ✓ 레이어 입력 x의 분포를 훈련과정으로 고정시킴으로 훈련 속도를 향상시키는 것을 기대해 테스트 함.

- ✓ **Whitening**기법을 적용시켜보였다. (데이터들의 분포를 평균0, 분산이 1이게 만들고 decorrelation시키는)
  - ✓ 데이터가 특정분포를 따르도록 강제할 수 있음. Internal Covariate Shift를 제거 가능하지만, 경사하강법을 통한 오차역전파가 제대로 이루어지지 않음.

  - ✓ 입력 벡터들로 정규화를 해서 해보려했지만, 연산량이 많아져서 비효율적이 됨.

- ✓ -> 새로운 방법을 찾아야함.
- ✓ (그래디언트 계산을 해야해서 미분이 가능한 연산, 파라미터를 업데이트할때마다 전체 데이터에 대한 연산을 요구하지 않아야 함.)

(3) Normalization via Mini-Batch Statistics

✓ 각 레이어의 입력에서 전체 whitening 비용은 많이 들고 모든 곳에서 미분할 수 없어 2가지 단순화를 만듬

1. 레이어 **입력 및 출력의 피쳐를 공동으로** whitenin하는 대신 평균을 0으로 하고, 분산을 1로 하여 각 **스칼라 feature를 독립적으로 정규화**
-> 이 방법은 시그모이드 활성함수를 거쳐 다음 레이어를 전달하게 되면 문제가 생길 수 있음.
(시그모이드 함수구간 [-1,1] 구간이 linear한 부분이라, 활성함수의 특징인 비선형성을 못가짐)
그래서 정규화를 진행하는 각 차원마다 새로운 2개의 파라미터인 γ (k) , β(k) 를 도입함. (scalining, shifting)
2. 배치를 사용한 한습에서 전체 데이터에 대한 평균과 분산 값을 구하는 것이 불가능.
배치에 대한 평균과 분산값을 통해 정규화를 진행.

이러한 방식은 전체 구간에 대해 미분이 가능함을 보이며, 정규화 과정이 정상적으로 진행됨에 따라 input에 대한 분포를 일정하게 유지할 수 있어, internal covariate shift문제 해결. 학습 속도 향상 가능.

(3) Normalization via Mini-Batch Statistics

- ✓ Inference과정에 대한 배치정규화
- ✓ Inference 과정에는 입력단위가 배치가 아닌 한 개의 입력 데이터이므로 평균과 분산을 어떻게 설정해야 하는지 정해야함.
- ✓ Training과정에서 사용한 배치들의 평균이 표본평균들을 사용해 모평균을 근사하여 사용하도록함.

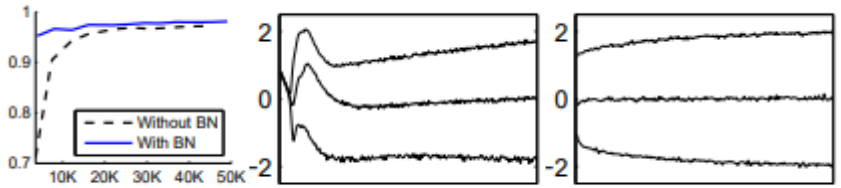
- ✓ 기존 딥러닝구조에서는 높은 학습률을 사용하게되면 그래디언트 값이 0에 가까워져서 학습이 이뤄지지 않는 문제가 있었지만, 배치 정규화를 쓰면 활성함수에서 중앙 부근으로 입력 분포를 바꿔주어 높은 학습률을 설정해도 괜찮음.

- ✓ Dropout은 일반적으로 과적합을 줄이기위해 사용하는데 배치정규화가 된 네트워크에서는 동일한 효과를 가질 수 있음. (특정 노드에 대한 activation을 제거하거나 감소시키는 효과)

(4) Experiments

✓ 내부 공변량 이동이 훈련에 미치는 영향과 배치 정규화가 훈련에 미치는 영향을 검증하기 위해 테스트

✓ MNIST데이터에서 숫자 클래스 예측 문제
✓ 배치 정규화를 했을 때 학습이 진행됨에 따라 더 안정적임.



(a) (b) Without BN (c) With BN

✓ 학습속도 또한 빠르다.
✓ BN 사용으로 train 단계 절반 이하의 정확도를 일치시킴.

(5) Conclusion

- ✓ 이러한 정규화는 네트워크를 교육하는데 최적화 방법에 의해 적절히 처리됨.
- ✓ 각각의 미니배치에 대해 정규화를 수행하고, 정규화 파라미터를 통해 그래디언트를 역전파한다.

- ✓ 이미지 분류 예시에서도 보면 학습속도를 높이고, Drop out을 제거하고, Batch Normalization이 제공하는 다른 수정을 적용해 좋은 성능을 가짐.

- ✓ 배치정규화의 목표는 훈련을 통해 활성화 값의 안정적인 분포를 달성하는 것이다.