PMLR 2015

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

Google Inc : Sergey Ioffe, Christian Szegedy

NIPS 2018

How Does Batch Normalization Help Optimization?

MIT : Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Mądry

고민수

Summary

Batch Normalization

레이어를 넘어갈 때 마다 ICS 가 커지는 것을 문제로 히든 레이어의 입력에 대한 정규화를 제안합니다.

- Batch 데이터 별 N(0,1) 정규화 진행
- Channel 별 2개의 파라미터를 추가하여 N(0,1) 정규화에서 발생할 수 있는 비선형 활성화 함수의 영향력 감소 문제 해결

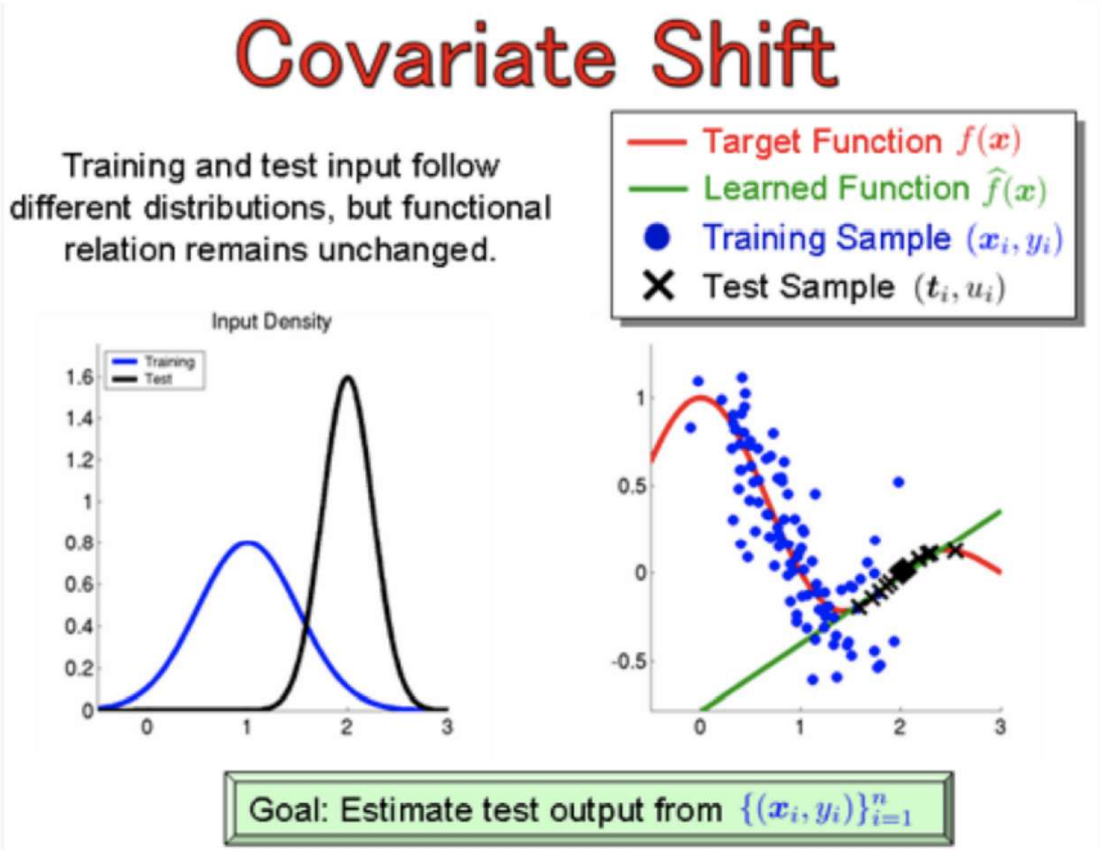
How Does Batch Normalization Help Optimization?

ICS는 Batch Normalization의 성능에 관계 없고 Smoothing 이 핵심이라고 설명합니다.

- 의도적으로 데이터에 노이즈를 넣어 엄청난 ICS를 발생시켰음에도 불구하고 BN이 일반 모델보다 강력한 성능을 보임으로 ICS는 BN의 성능과 관계없음을 증명합니다.
- BN의 핵심은 Smoothing이고 이를 통해 기울기의 예측성(predictiveness), Lipschitzness 가 좋아져 안정적인 학습이 가능함을 증명합니다.

ICS : Internal Covariate Shift

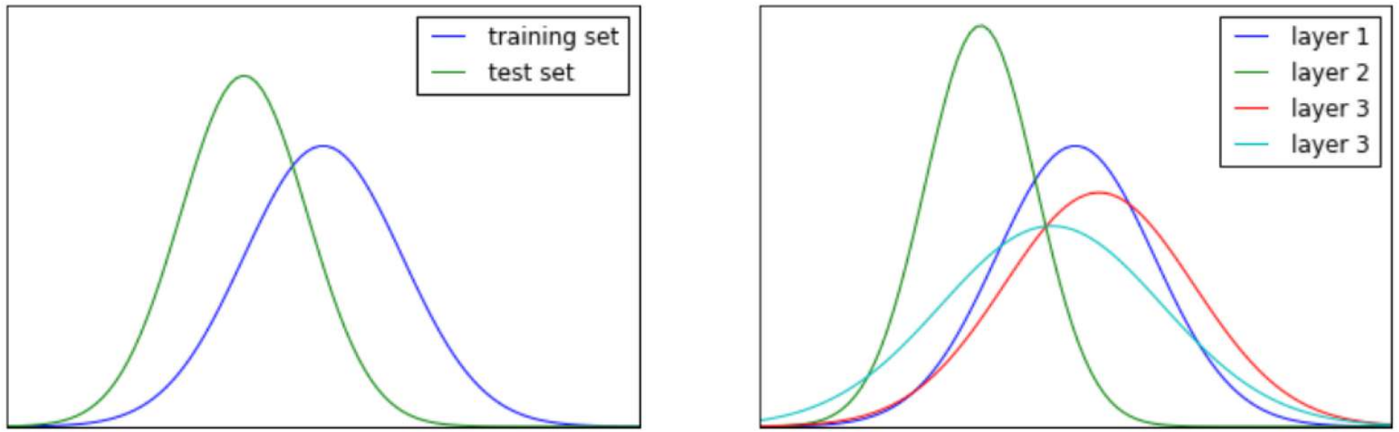
공변량 변화(Covariate Shift) : 학습시기와 테스트 시기에 입력 데이터의 분포가 변화하는 경우



Dataset Shift, The MIT Press

ICS : Internal Covariate Shift

Internal Covariate Shift : Covariate Shift가 네트워크 내부에서 발생하는 현상



(a) Covariate shift (b) Internal covariate shift

Figure 3.1: Covariate shift vs. internal covariate shift

파라미터가 업데이트됨에 따라 Hidden layer들이 입력 분포가 변경됩니다. 뒤의 레이어 입장에서 매 스텝마다 입력 분포가 바뀌는 것과 동일하며 레이어가 깊어질수록 심화될 수 있습니다.

The Effect of Batch Normalization on Deep Convolutional Neural Networks(2016)-FABIAN SCHILLING

Batch Nomalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

입력(Input) : Batch 데이터
입력 차원 수 X 학습파라미터 2개

출력(Output) : BN이 적용된 Batch 데이터

배치 평균

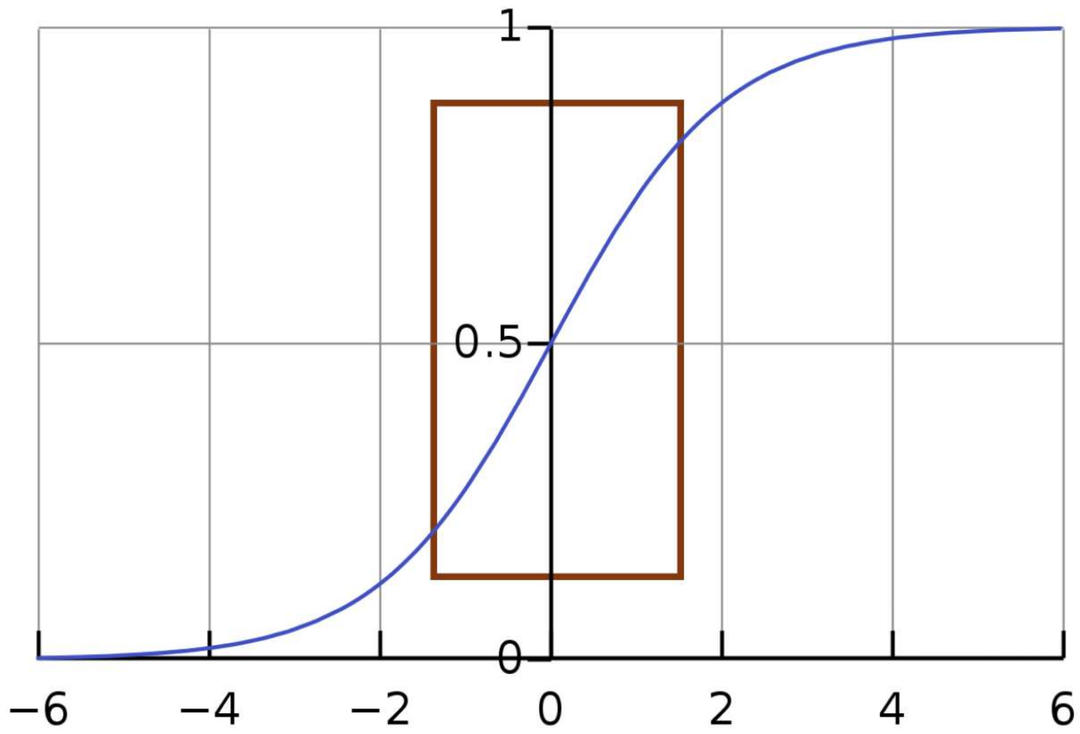
배치 분산

정규화, 앱실론은 아주작은 상수(오류방지)

스케일 조절 및 이동

Batch Nomalization

스케일 조절 및 이동의 이유 $\mathrm{z} = g(W\mathrm{u} + \mathrm{b}) \longrightarrow \mathrm{z} = g(\mathrm{BN}(W\mathrm{u}))$



N(0,1)로 정규화를 했을때 대부분의 입력에 대해 비선형 활성화 함수의 영향력이 감소하고 수렴, 학습이 느려질 수 있습니다.

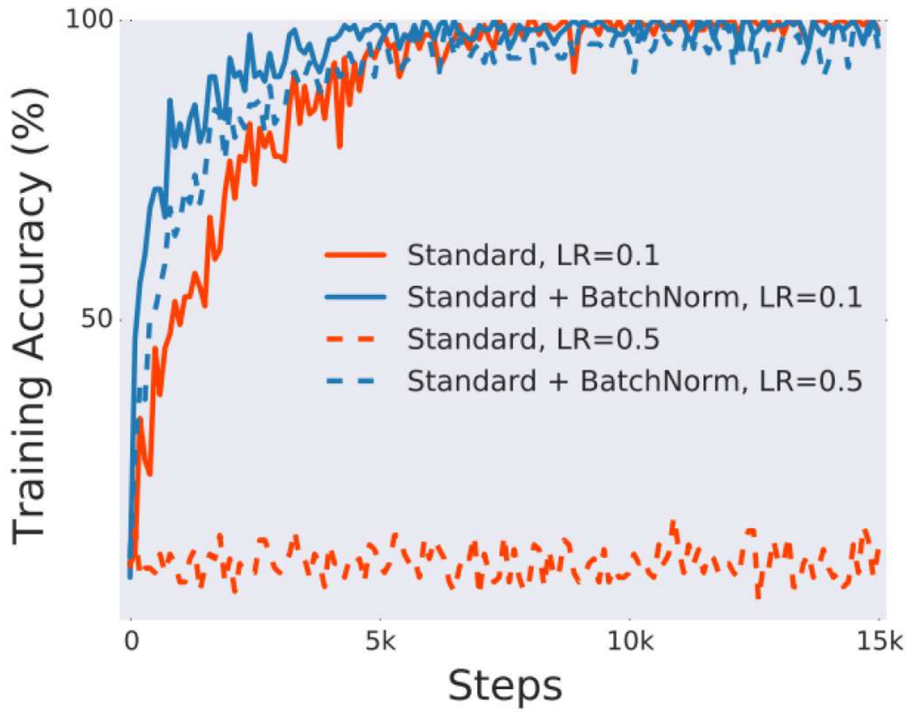
↓

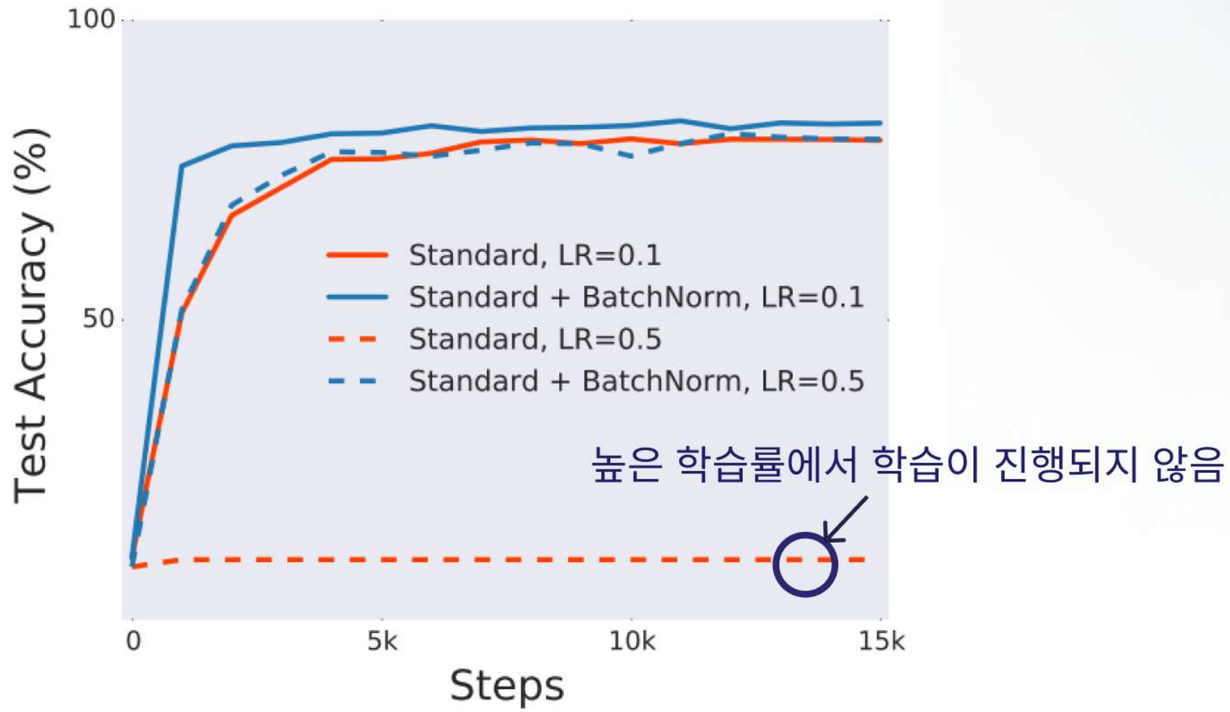
W, u 를 통해 정규화된 값을 조정하고 이동합니다. 또한 정규화에 사용된 통계가 역전파에 참여할 수 있도록 합니다.

Wikipidia - Sigmoid

Batch Nomalization 의 장점

학습을 위한 하이퍼 파라미터 설정에서 관대해지고, 수렴속도가 월등히 빠릅니다





How Does Batch Normalization Help Otimization?(NIPS 2018)

NIPS 2018

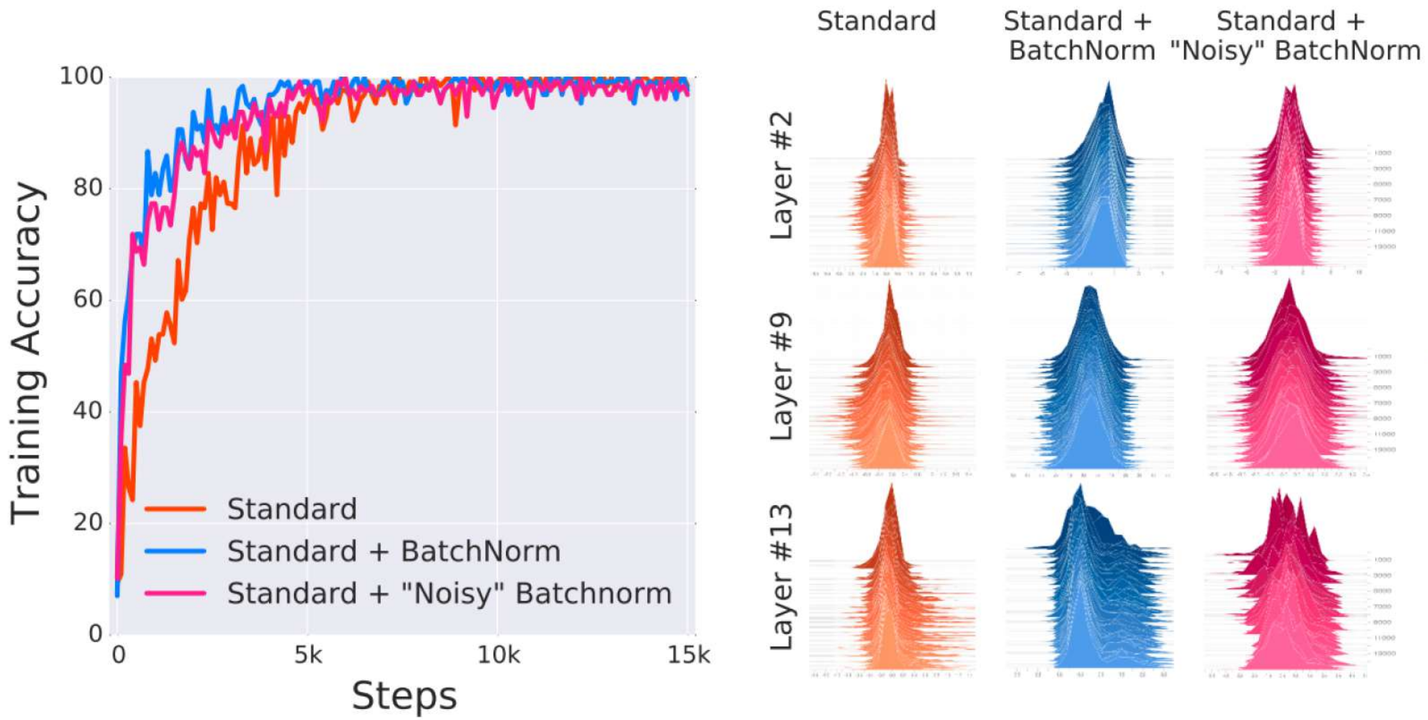
How Does Batch Normalization Help Optimization?

MIT : Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Mądry

본 논문에서 Batch Normalization의 효과는 ICS의 감소와 연관이 없다는 주장이 제기됩니다

Compare BN & “Noisy”BN

Noise를 임의로 추가하여 ICS를 크게 발생 시킨 데이터와 비교



ICS를 크게 발생시킨 데이터에 BN을 적용한 모델이 BN을 사용하지 않은 모델보다 수렴속도가 월등히 빠르며, 기존의 BN모델과 성능이 거의 유사한 것을 통해 ICS는 BN의 성능과 무관함을 증명합니다.

Figure 2: Connections between distributional stability and BatchNorm performance

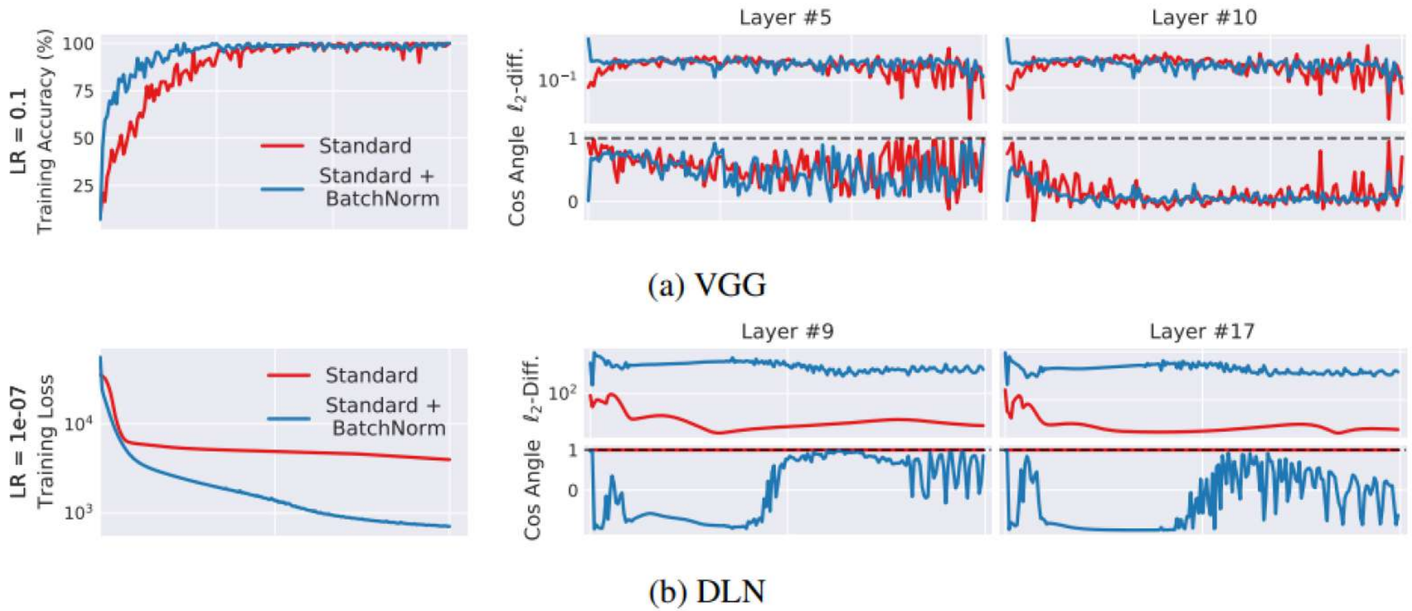
Define ICS diff

동일한 입력에서 K번째 레이어의 모든 앞쪽 파라미터를 update 한 것과 하지 않았을 때 Gradient를 비교

$$G_{t,i} = \nabla_{W_i^{(t)}} \mathcal{L}(W_1^{(t)}, \ldots, W_k^{(t)}; x^{(t)}, y^{(t)})$$

$$G'_{t,i} = \nabla_{W_i^{(t)}} \mathcal{L}(W_1^{(t+1)}, \ldots, W_{i-1}^{(t+1)}, W_i^{(t)}, W_{i+1}^{(t)}, \ldots, W_k^{(t)}; x^{(t)}, y^{(t)}).$$

ICS diff = $||G_{t,i} - G'_{t,i}||_2$

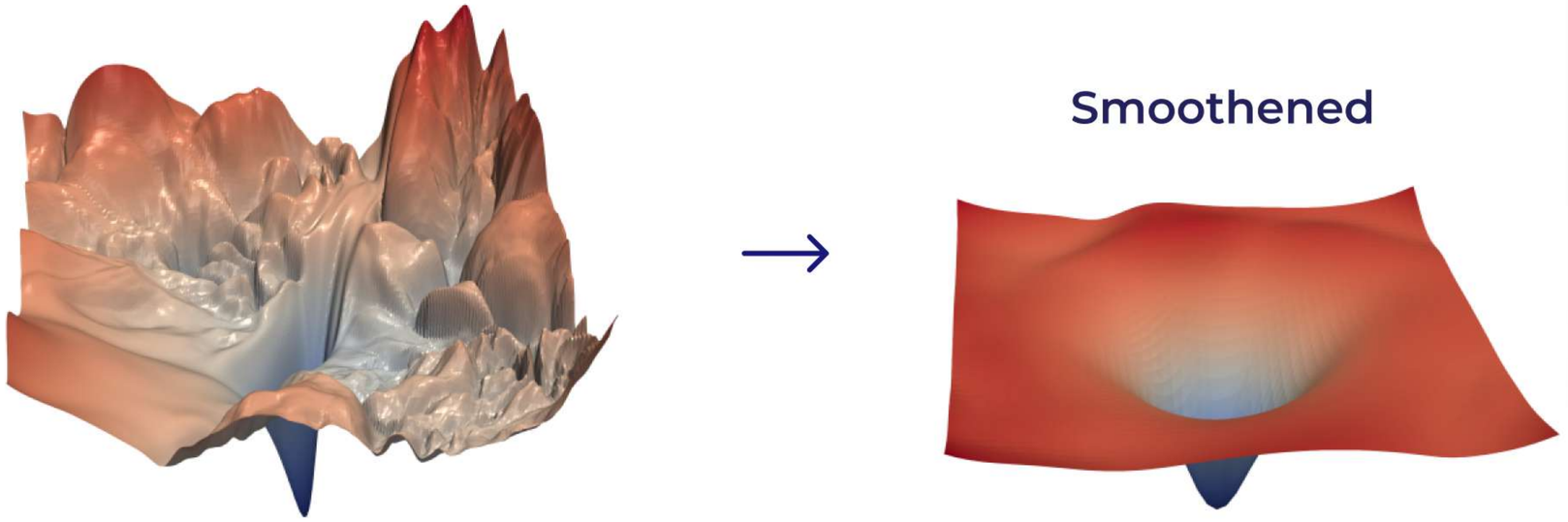


→

BN을 사용해도 ICS에 큰 차이가 없고 오히려 증가하는 경우가 많은 것을 볼 때 최적화 관점에서 BN은 내부공변량이동(ICS)을 줄이지 못함을 시사합니다.

Smoothing

BN이 대부분 모델에서 좋은 효과를 보이는 것은 Loss Landscape를 Smoothing 때문이라고 제시합니다.



- Smoothing은 기울기를 보다 안정적이고 예측가능하게 만들어줍니다. 즉 향상된 Lipschitzness는 계산된 Gradient 방향으로 더 큰 단계를 밟을 때 상당히 정확한 추정치를 유지해줍니다.
- 이를 통해 Gradient 소실 또는 폭발의 환경에서 위험없이 큰 단계의 학습을 수행하도록 합니다.

Visualizing the Loss Landscape of Neural Nets (NIPS 2018)

Thank you

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

Google Inc : Sergey Ioffe, Christian Szegedy

How Does Batch Normalization Help Optimization?

MIT : Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Mądry