한국과학기술정보연구원
Korea Institute of Science and Technology Information
www.kisti.re.kr
UST

NLP Seminar

Batch Normalization & Layer Normalization

2022. 04. 29

KISTI - UST IKJE CHOI

CONTENTS

Batch Normalization

Batch Normalization

- Title : Batch normalization: Accelerating deep network training by reducing internal covariate shift

- Google Scholar

  Batch normalization: Accelerating deep network training by reducing internal covariate shift

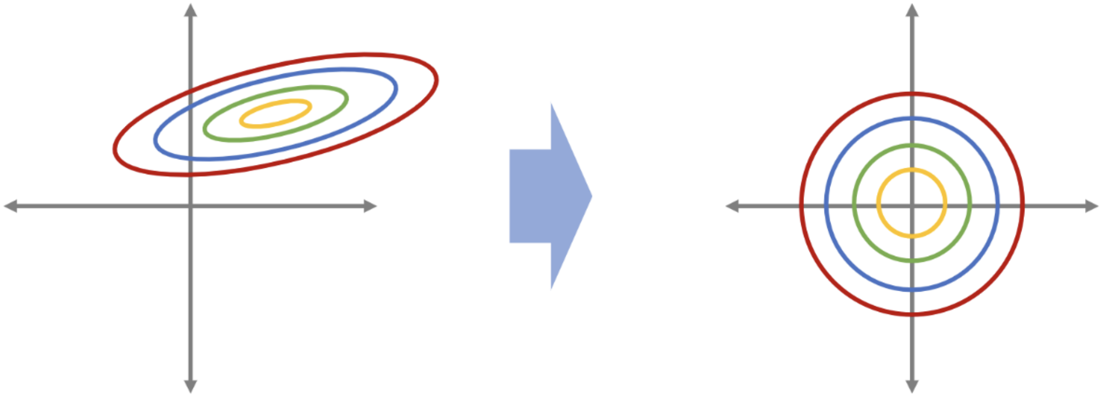
  S Ioffe, C Szegedy - International conference on machine …, 2015 - proceedings.mlr.press

  Abstract Training Deep Neural Networks is complicated by the fact that the distribution of each layer's inputs changes during training, as the parameters of the previous layers …

  Save Cite Cited by 36144 Related articles All 40 versions

Batch Normalization

- ➢ Batch
  - ✓ Epoch : one pass over the full training set
  - ✓ Batch : use all data to compute the gradient during one iteration.
  - ✓ Mini-batch : take a subset of all data during one iteration.
- ➢ Normalization
  - ✓ Gets rid of a variety of irregularities that can make it more difficult to interpret the data.



What is the advantages of using Batch Normalization?

- Reducing <u>Training Time</u>
- Reducing significant changes in <u>Weight Initialization</u>
- Provide <u>Regularization</u> effect of the model

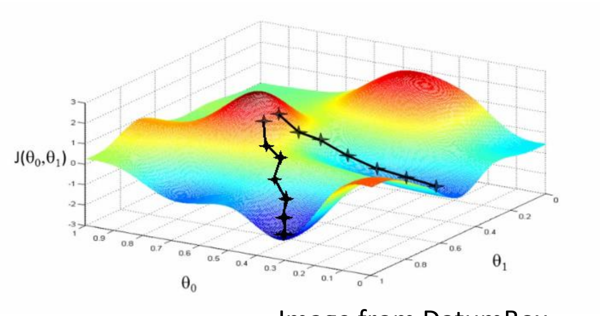
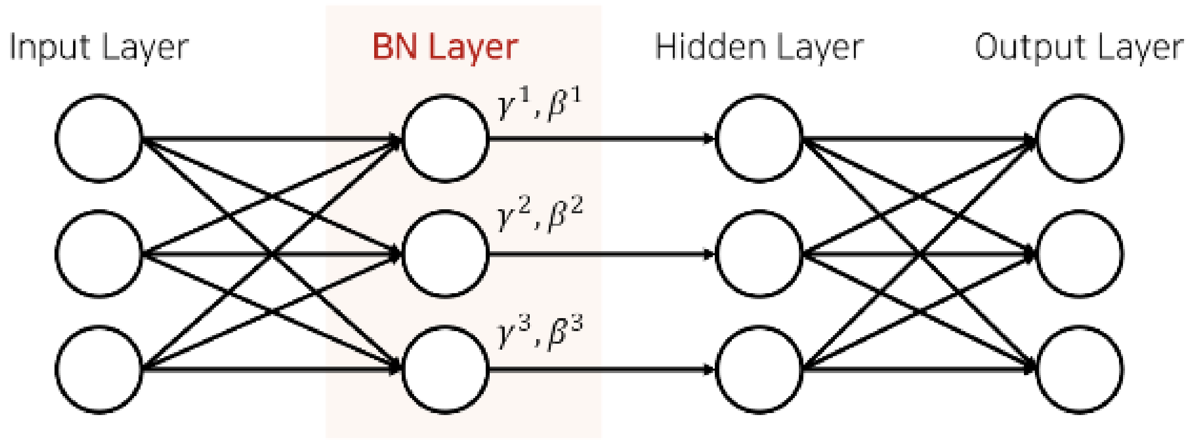
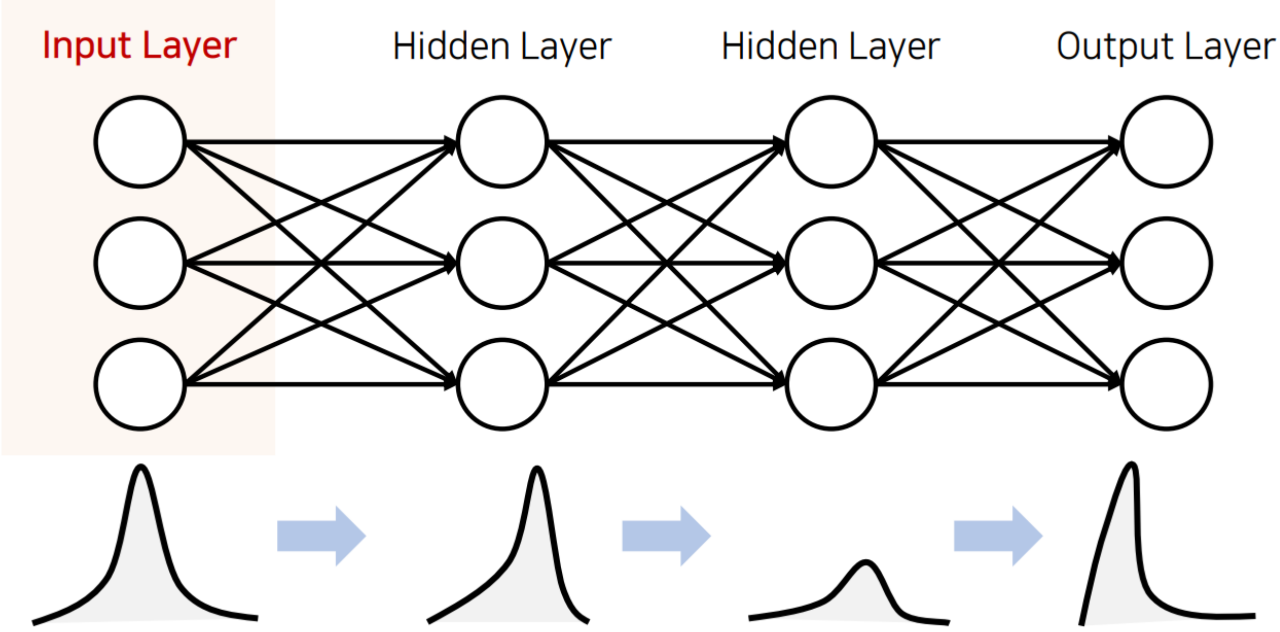


Image from DatumBox



Motivation

- Scaling(Normalizing) data then it gives better accuracy and faster speed.
- If this is applied to hidden Layer then It might give same advantages.



Normalization Equation

```python
x1 = np.asarray([33, 72, 40, 104, 52, 56, 89, 24, 52, 73])
x2 = np.asarray([9, 8, 7, 10, 5, 8, 7, 9, 8, 7])
```

- Mean

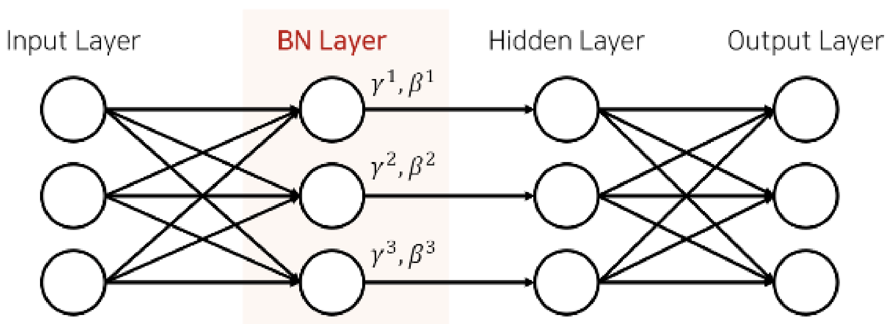
$$\mu_{Batch} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

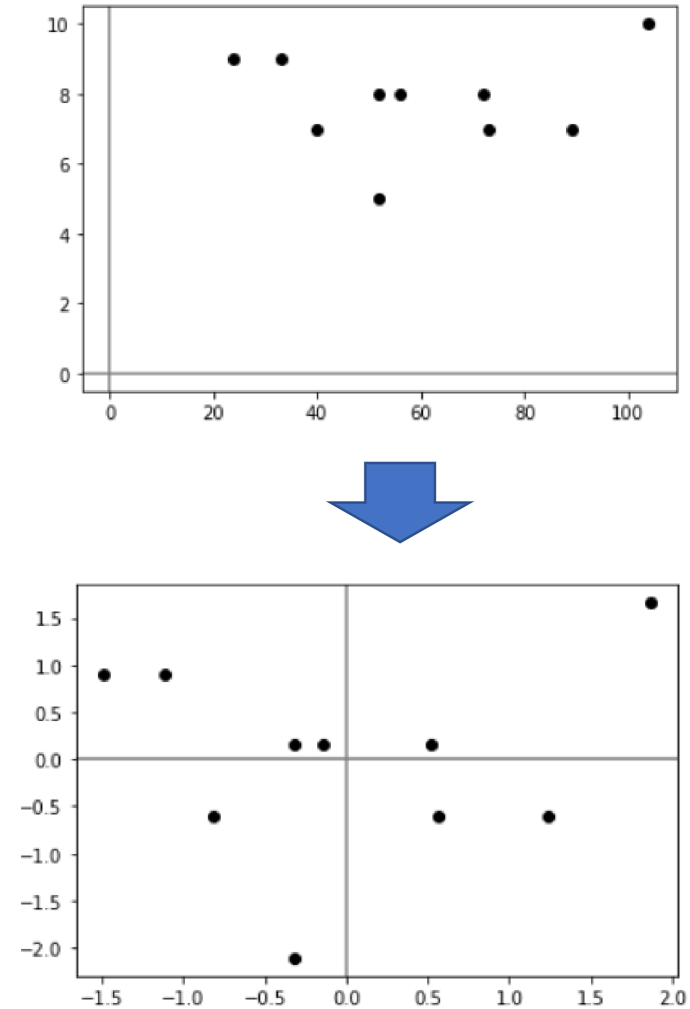
- Variance

$$\sigma_{Batch}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{Batch})^2$$

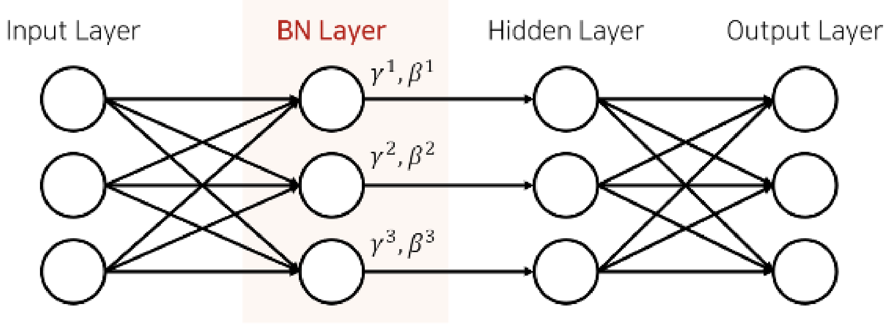
- Normalization

$$\widehat{x_i} \leftarrow \frac{x_i - \mu_{Batch}}{\sqrt{\sigma_{Batch}^2 \boxed{+ \epsilon}}}$$





Batch Normalization

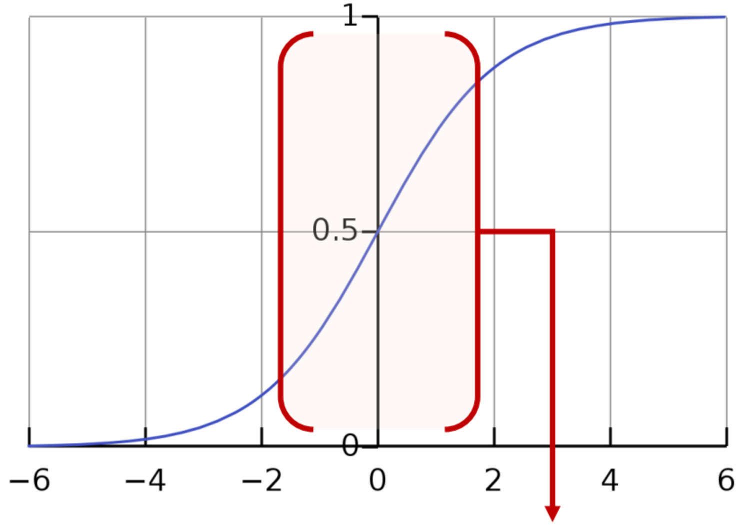


- Normalization $\hat{x}_i \leftarrow \dfrac{x_i - \mu_{Batch}}{\sqrt{\sigma^2_{Batch} + \epsilon}}$

Why two parameters (γ, β) learning?

- Equation $y_i \leftarrow \gamma\hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$

- Loosing Non-linearity (example : Sigmoid)
  - ✓ Linearity -> Non-linearity by learning (γ, β)



Almost Linear in Sigmoid

Batch Normalization

Batch Normalization for Train

- Mean

$$\mu_{Batch} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Variance

$$\sigma_{Batch}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{Batch})^2$$

- Normalization

$$\widehat{x_i} \leftarrow \frac{x_i - \mu_{Batch}}{\sqrt{\sigma_{Batch}^2 + \epsilon}}$$

$$y_i \leftarrow \gamma\widehat{x_i} + \beta \equiv BN_{\gamma,\beta}(x_i)$$

Batch Normalization for Test

- No mini batch at test
- But we still need to apply BN at test
- Then How?
- Mean

$$E[x] \leftarrow E_{Batch}[\mu_{Batch}]$$

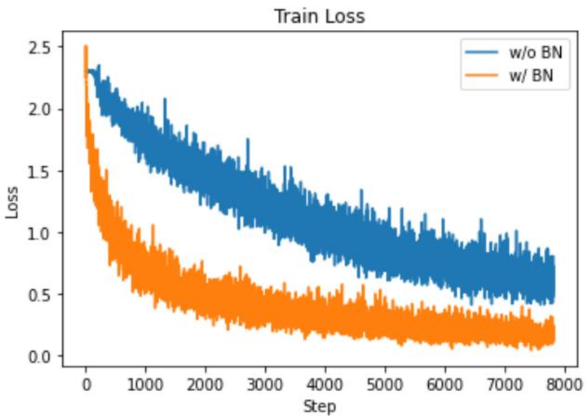
- Variance $$Var[x] \leftarrow \frac{m}{m-1}E_{Batch}[\sigma_{Batch}^2]$$

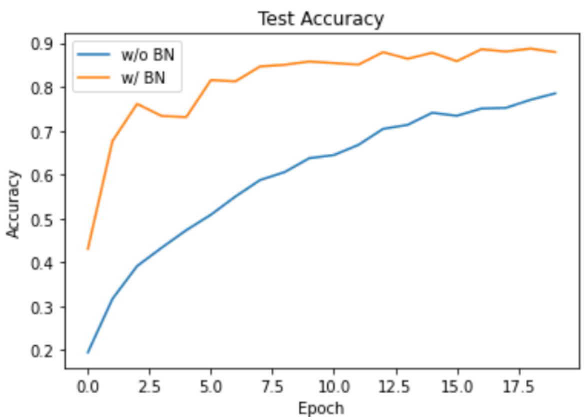
$\mathrm{y} = BN_{\gamma,\beta}(x)$ with $\mathrm{y} = \frac{\gamma}{\sqrt{Var[x]+\epsilon}} \cdot x + \left(\beta - \frac{\gamma E[x]}{\sqrt{Var[x]+\epsilon}}\right)$

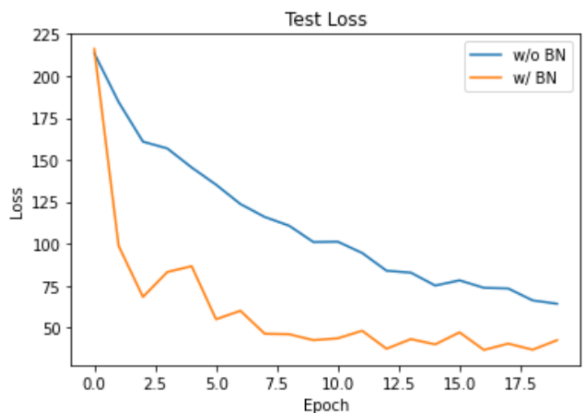
Batch Normalization

What is the advantages of using Batch Normalization?

- Reducing Training Time (14 times)
- Reducing significant changes in Weight Initialization
- Provide Regularization effect of the model







Batch Normalization

➢ Google Scholar

Batch normalization: Accelerating deep network training by reducing internal covariate shift

S Ioffe, C Szegedy - International conference on machine …, 2015 - proceedings.mlr.press

Abstract Training Deep Neural Networks is complicated by the fact that the distribution of each layer's inputs changes during training, as the parameters of the previous layers …

☆ Save 99 Cite Cited by 36144 Related articles All 40 versions

How does **batch normalization** help optimization?

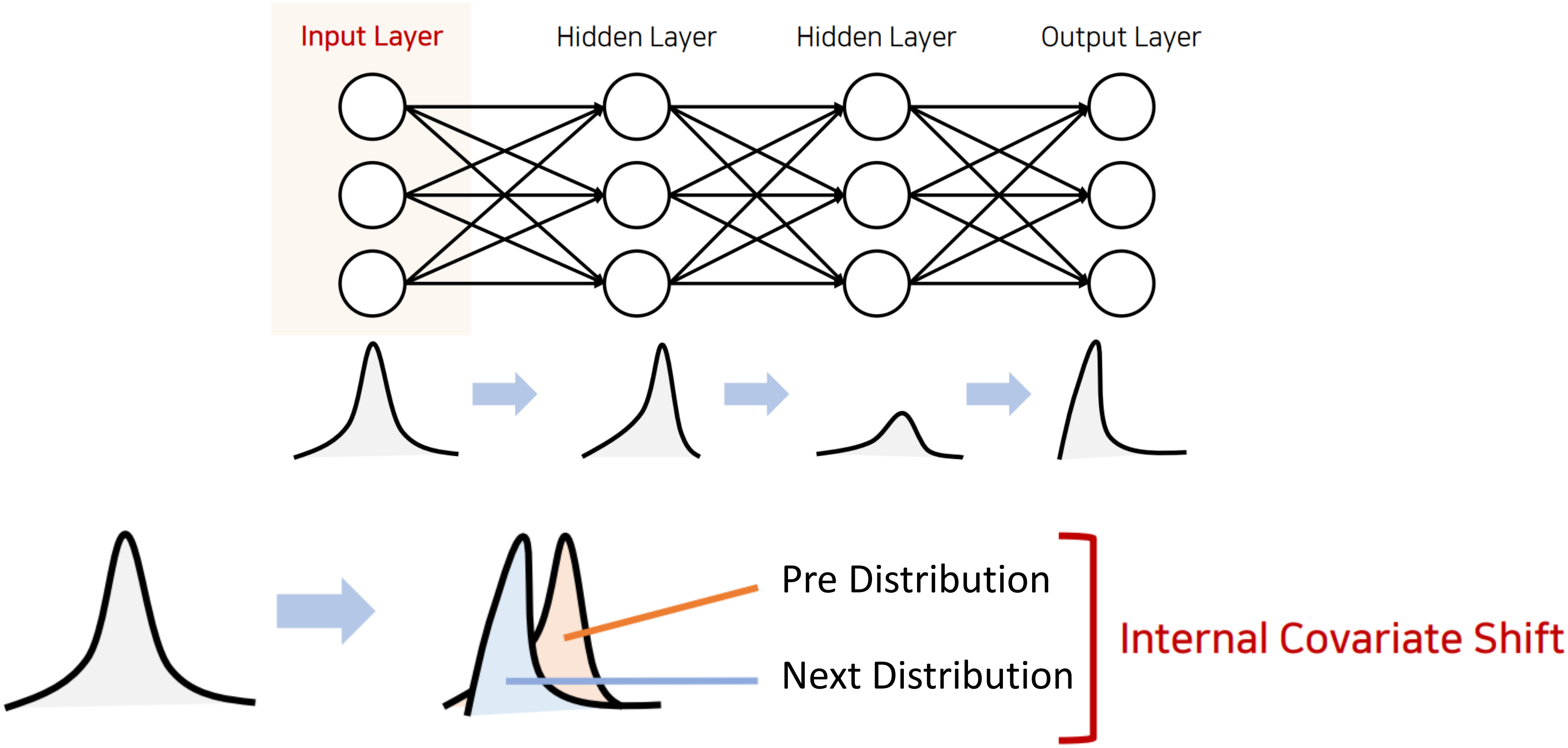
S Santurkar, D Tsipras, A Ilyas… - Advances in neural …, 2018 - proceedings.neurips.cc

… We show that **batch normalization** causes this landscape to be more well-behaved, inducing favourable properties in Lipschitz-continuity, and predictability of the gradients. We then …

☆ Save 99 Cite Cited by 1034 Related articles All 15 versions

Internal Covariate Shift

➢ Original Paper



Batch Internal Covariate Shift

- Later Paper

  How does **batch normalization** help optimization?

  S Santurkar, D Tsipras, A Ilyas… - Advances in neural …, 2018 - proceedings.neurips.cc

  … We show that **batch normalization** causes this landscape to be more well-behaved, inducing favourable properties in Lipschitz-continuity, and predictability of the gradients. We then …

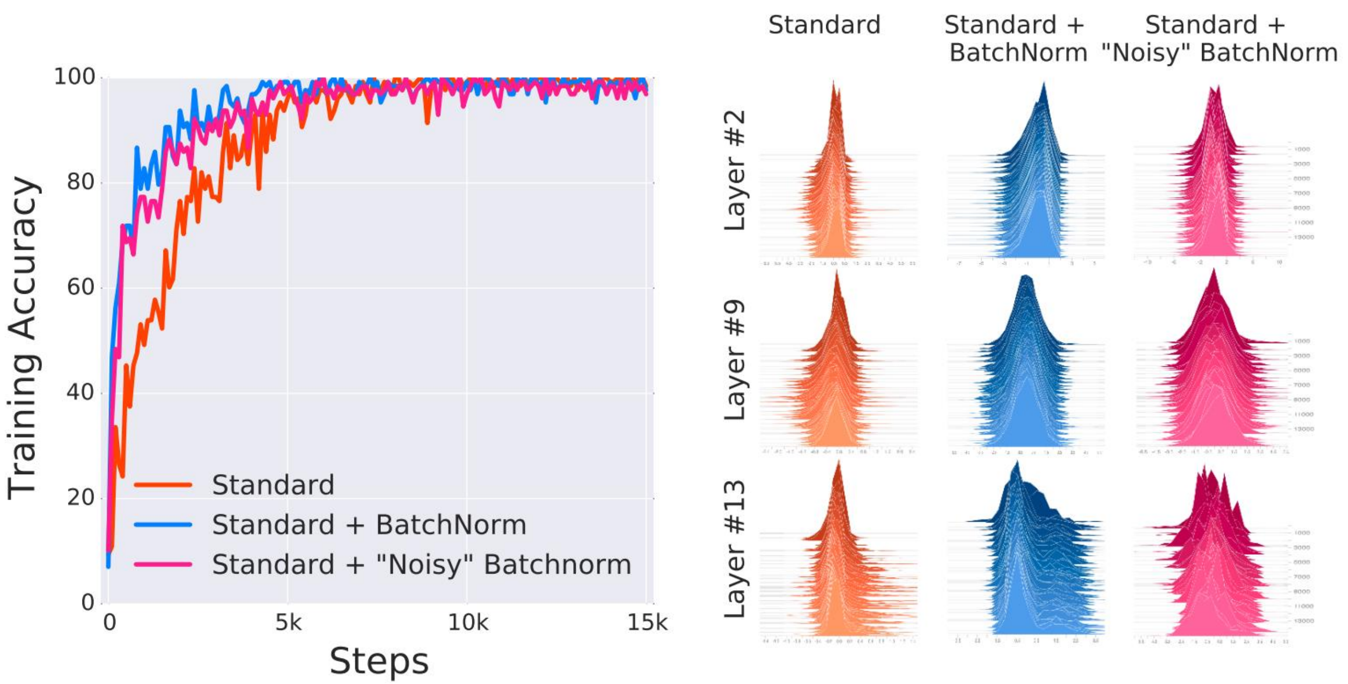
  Save Cite Cited by 1034 Related articles All 15 versions

- This paper agrees about all technical advantages of BN

- But it does not relate any to Internal Covariate Shift

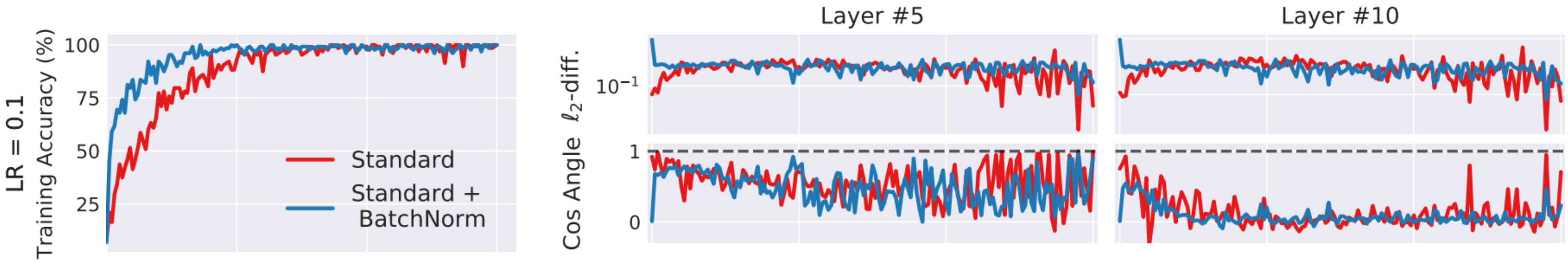
Internal Covariate Shift

- Intentionally add noise and test accuracy -> increase internal covariate shift
- Still BN with noise is better than Standard



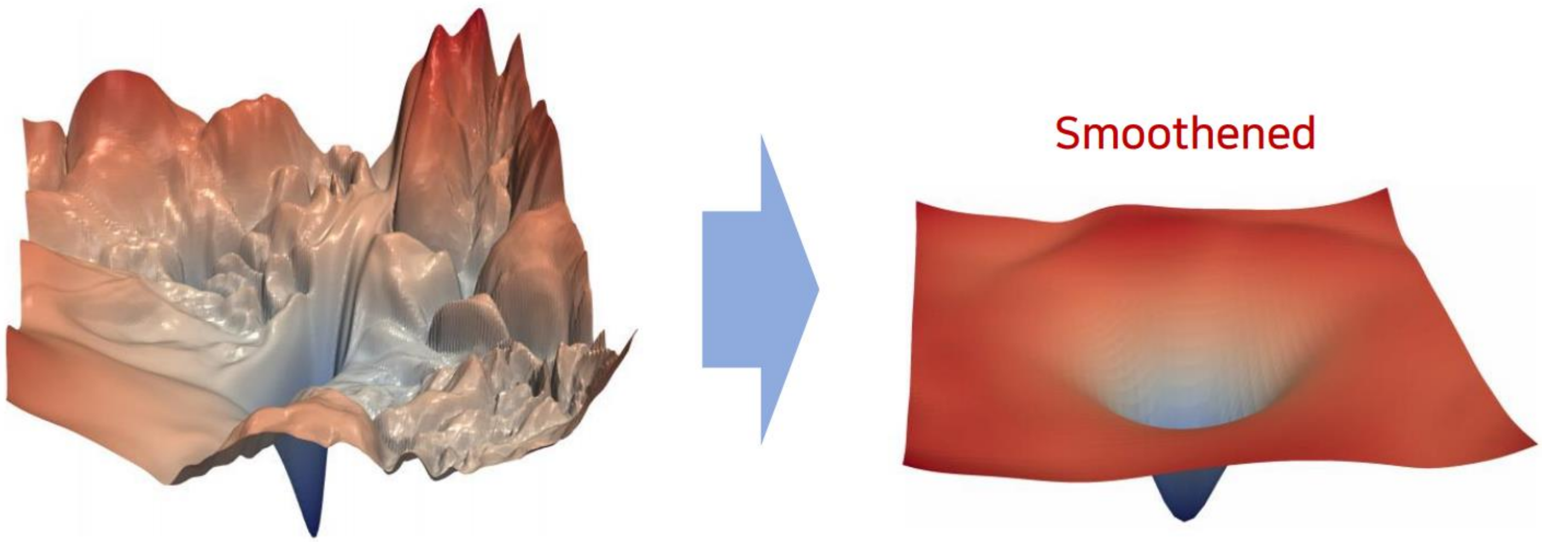
Internal Covariate Shift calculation

- Every Weight update and calculate similarity
- Plot between with BN and without BN



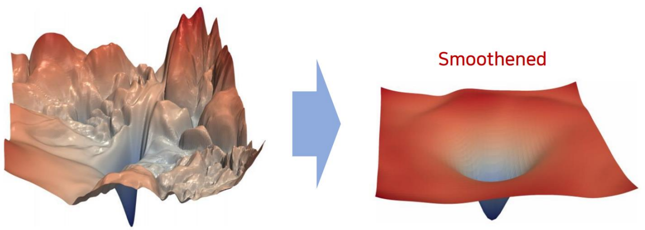
What is reason that BN works well?

- Smoothing Effect!

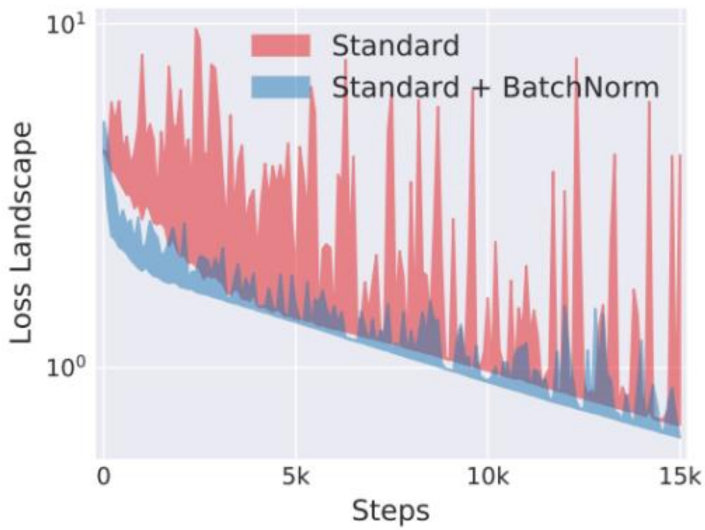


Batch Internal Covariate Shift

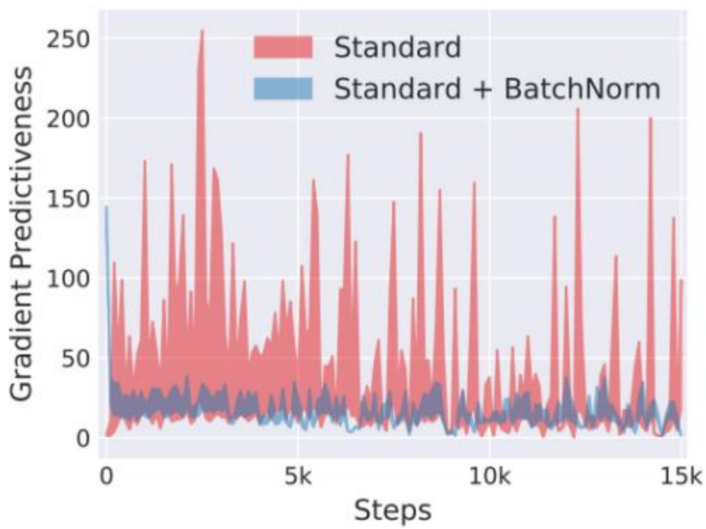
- Every Steps calculating Loss and Gradient
  - ✓ Big differences imply less reliable gradients
  - ✓ Large Fluctuation make optimization hard



Variation in Loss $(L(W))$



Change in Gradient $(\nabla_W L(W))$



Layer Normalization

- Google Scholar

Layer normalization

JL Ba, JR Kiros, GE Hinton - arXiv preprint arXiv:1607.06450, 2016 - arxiv.org

… , we transpose batch **normalization** into **layer normalization** by computing the mean and variance used for **normalization** from all of the summed inputs to the neurons in a **layer** on a …

Save Cite Cited by 4951 Related articles All 7 versions

Layer Normalization

- Batch Normalization
  - ✓ Hard to use with Sequence data as Sequence data has varying length
  - ✓ Batch means Mini-Batch in Batch Normalization
    It is hard to parallelization as it BN has dependency on its batches.

- Layer Normalization
  - ✓ LN remove dependency as it applies normalization based on layer.

Layer Normalization

Batch Normalization

- Mean

$$\mu_{Batch} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Variance

$$\sigma_{Batch}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{Batch})^2$$

- Normalization

$$\widehat{x_i} \leftarrow \frac{x_i - \mu_{Batch}}{\sqrt{\sigma_{Batch}^2 + \epsilon}}$$

Layer Normalization

- Mean

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l$$

- Variance

$$\sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H} \left(a_i^l - \mu^l\right)^2}$$

- Normalization

$$\mathbf{h}^t = f\left[\frac{\mathbf{g}}{\sigma^t} \odot \left(\mathbf{a}^t - \mu^t\right) + \mathbf{b}\right]$$

Layer Normalization

- Normalization $$\mathbf{h}^t = f\left[\frac{\mathbf{g}}{\sigma^t} \odot \left(\mathbf{a}^t - \mu^t\right) + \mathbf{b}\right]$$

- Weight re-scaling and re-centering

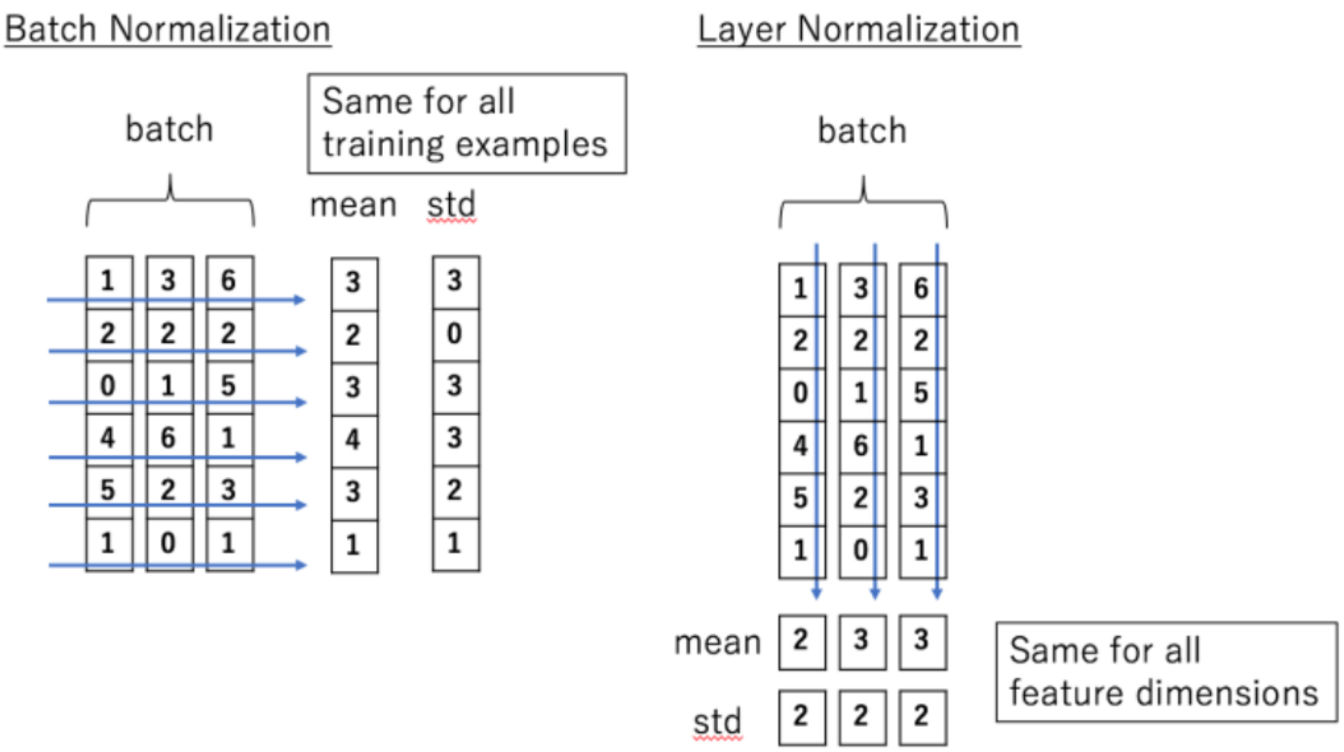
$$\begin{aligned}\mathbf{h}' =& f\left(\frac{\mathbf{g}}{\sigma'}\left(W'\mathbf{x} - \mu'\right) + \mathbf{b}\right) = f\left(\frac{\mathbf{g}}{\sigma'}\left(\left(\delta W + \mathbf{1}\boldsymbol{\gamma}^\top\right)\mathbf{x} - \mu'\right) + \mathbf{b}\right) \\ =& f\left(\frac{\mathbf{g}}{\sigma}\left(W\mathbf{x} - \mu\right) + \mathbf{b}\right) = \mathbf{h}.\end{aligned}$$

Where bias for b, gain for g

- Easy to see re-scaling individual data points does not change the model's prediction under layer normalization

Layer Normalization

- Layer Normalization is effective at longer sequences data
- Suitable with RNN but not CNN (BN is recommended)



Reference

Paper

- 2015, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift” by Sergey Ioffe
- 2018, “How Does Batch Normalization Help Optimization?” by Shibani Santurkar
- 2016, “Layer Normalization” by Jimmy Lei Ba

Extra

- https://github.com/ndb796/Deep-Learning-Paper-Review-and-Practice

THANK YOU