Machine Learning (CE 40477)

Fall 2024

Ali Sharifi-Zarchi

CE Department
Sharif University of Technology

October 8, 2024

1. Batch Normalization

2. References

1. Batch Normalization
   - Batch Normalization introduction
   - Why Batch Normalization?
   - How Batch Normalization Works
   - Batch Normalization Pros & Cons
   - Batch Normalization in Practice
   - Closing Takeaways on Batch Normalization
2. References

What is Batch Normalization Concept?

- Batch Normalization main purpose: **Smoothing the optimization space**
- Batch Normalization Opt.: Normalizing activations in a network.

Smoothing the optimization space

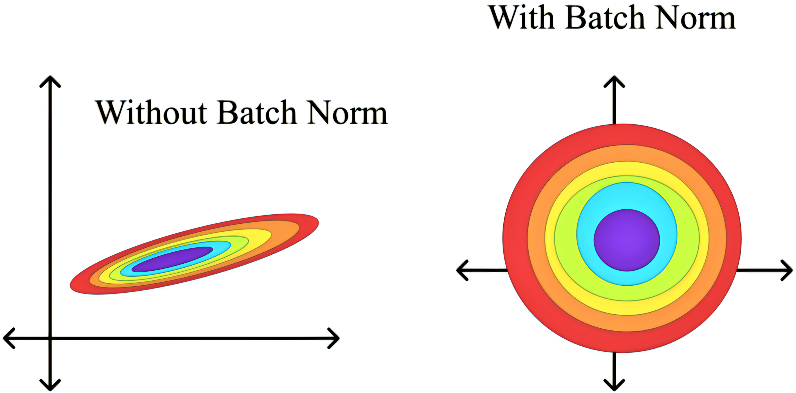


Figure 1: Using Batch Normalization causes the optimization space to become smoother. Source

1. Batch Normalization
   - Batch Normalization introduction
   - Why Batch Normalization?
   - How Batch Normalization Works
   - Batch Normalization Pros & Cons
   - Batch Normalization in Practice
   - Closing Takeaways on Batch Normalization
2. References

Why Batch Normalization?

Problem: Internal Covariate Shift

- **Definition:** During training, as the parameters of earlier layers change, the distribution of inputs to deeper layers shifts, which slows down learning.
- **Impact:** This shift makes the network slower to train and requires careful tuning of the learning rate.
- **Example:** Imagine training a deep network where the input scale and distribution to each layer are constantly changing, making convergence harder to achieve.

Why Batch Normalization?

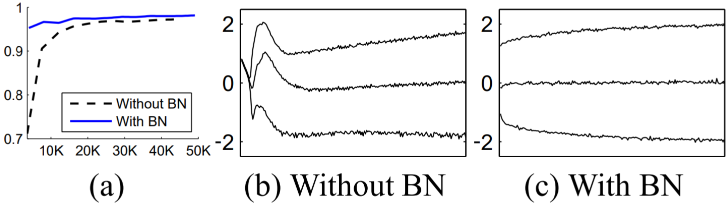
Batch Normalization Solution

- **Goal:** Normalize the inputs to each layer so that the mean is close to 0 and the variance is close to 1.
- **How it helps:** This stabilization of the learning process allows for higher learning rates and often leads to faster convergence.
- **Additional Benefits:**
  - Reduces sensitivity to weight initialization.
  - Mitigates vanishing and exploding gradient problems in deep neural networks.

Magic Effect of Batch Normalization

Magic Effect!

- Batch Normalization helps the network train faster and achieve higher accuracy.
- Batch Normalization **makes the distribution more stable** and reduces the internal covariate shift.



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

Figure 2: (a) The test accuracy of the MNIST network trained with and without Batch Normalization, vs the number of training steps. (b, c) The evolution of input distributions to a typical sigmoid, over the course of training, shown as 15, 50, 85th percentiles [1].

1 Batch Normalization

- Batch Normalization introduction
- Why Batch Normalization?
- How Batch Normalization Works
- Batch Normalization Pros & Cons
- Batch Normalization in Practice
- Closing Takeaways on Batch Normalization

2 References

How Batch Normalization Works

Process Overview

- For each mini-batch during training, batch normalization normalizes the inputs to a layer by adjusting their mean and variance.

How Batch Normalization Works

Steps in Batch Normalization

1. **Compute the Mean and Variance**
   For a given mini-batch, compute the mean μ_B and variance σ_B^2 of the inputs:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$$

How Batch Normalization Works

Steps in Batch Normalization

1. **Compute the Mean and Variance**
   For a given mini-batch, compute the mean μ_B and variance σ_B^2 of the inputs:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2$$

2. **Normalize the Inputs**
   Subtract the mean and divide by the standard deviation to get normalized activations:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

   where ϵ is a small constant added for numerical stability.

How Batch Normalization Works (Continued)

Steps in Batch Normalization

3. **Scale and Shift**

   After normalization, introduce learnable parameters γ and β that allow the model to scale and shift the normalized output:

   $$y_i = \gamma \hat{x}_i + \beta$$

   This ensures that the model can recover the original data distribution if needed.

How Batch Normalization Works

Inference Mode: Hint!!!

- During inference (when predicting new data), batch statistics (mean and variance) are replaced with moving averages collected during training.

Effect of Batch Normalization on Gradients

The main benefit of batch normalization is that it reduces the dependency of the gradient on the scale of the input and parameters:

$$\frac{\partial \mathscr{L}}{\partial x} = \frac{\partial \mathscr{L}}{\partial y} \cdot \frac{\partial y}{\partial \hat{x}} \cdot \frac{\partial \hat{x}}{\partial x} \tag{1}$$

Where:

- $\frac{\partial y}{\partial \hat{x}} = \gamma$
- $\frac{\partial \hat{x}}{\partial x} = \frac{1}{\sqrt{\sigma_{\mathscr{B}}^2 + \epsilon}}$

Thus, the gradient becomes:

$$\frac{\partial \mathscr{L}}{\partial x} = \frac{\gamma}{\sqrt{\sigma_{\mathscr{B}}^2 + \epsilon}} \cdot \frac{\partial \mathscr{L}}{\partial y} \tag{2}$$

How Batch Normalization Smooth the Cost Surface

The smoothing effect of batch normalization can be understood by observing how it constrains the gradient magnitudes. The expression shows that:

$$\frac{\partial \mathscr{L}}{\partial x} \text{ is scaled by } \frac{1}{\sqrt{\sigma_{\mathscr{B}}^2 + \epsilon}} \tag{3}$$

This consistent scaling leads to a smoother loss surface because:

- It stabilizes the gradient flow, ensuring controlled optimization step sizes.
- Reduces the risk of large oscillations or abrupt changes in the loss landscape.
- Makes the optimization process less likely to be trapped in local minima or saddle points.

1 Batch Normalization

Batch Normalization introduction

Why Batch Normalization?

How Batch Normalization Works

Batch Normalization Pros & Cons

Batch Normalization in Practice

Closing Takeaways on Batch Normalization

2 References

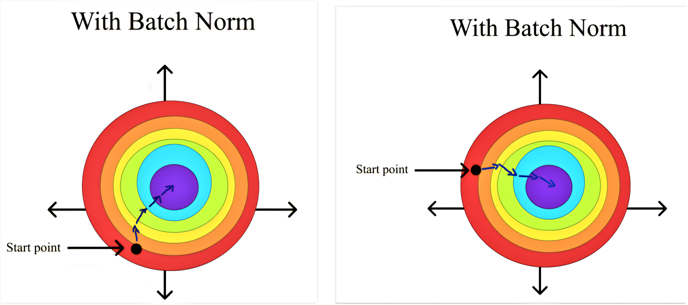
Batch Normalization Pros

Pros

- **Faster Convergence:** Empirical results support that models with batch normalization converge faster and achieve higher accuracy, even with higher learning rates.
- **Reduced Sensitivity to Weight Initialization:** Helps mitigate the dependency on careful weight initialization.
- **Acts as Regularization:** Batch normalization can help reduce overfitting.
- **Reduces Vanishing/Exploding Gradients:** Helps maintain stable gradients throughout deep networks.

Batch Normalization Pros

Why Using Batch Normalization Reduces Sensitivity to Weight Initialization?



(a) (b)

Figure 3: Start point doesn't matter! Source

Batch Normalization Pros

Why Using Batch Normalization Reduces Sensitivity to Weight Initialization?

- Batch normalization decreases the importance of initial weights because makes optimization space smoother,
  so doesn't matter where you start, you will get to the minimal point more or less with the same quantity of iterations in every start point.

Batch Normalization Cons

Cons

- **Batch Size Sensitivity:** Performance can depend on batch size, and very small batches may not provide stable statistics.
- **Computational Overhead:** Adds extra computation during training.
- **Behavior During Inference:** The shift from batch statistics to moving averages during inference may lead to slight discrepancies.

1 Batch Normalization

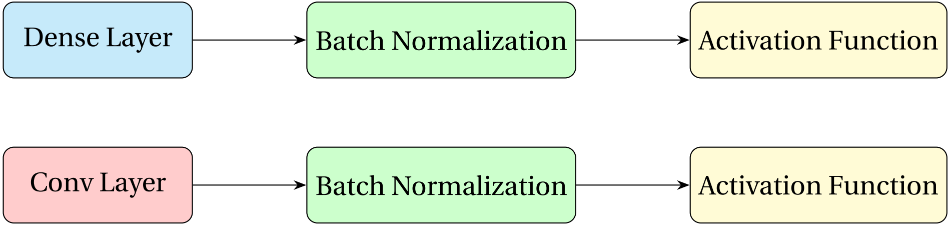
- Batch Normalization introduction
- Why Batch Normalization?
- How Batch Normalization Works
- Batch Normalization Pros & Cons
- Batch Normalization in Practice
- Closing Takeaways on Batch Normalization

2 References

Batch Normalization in Practice

Where to Apply

- **Typical Location:** Apply after the linear transformation (e.g., after a dense or convolutional layer) but before the activation function.
- **Layer Placement:**



1 Batch Normalization

- Batch Normalization introduction
- Why Batch Normalization?
- How Batch Normalization Works
- Batch Normalization Pros & Cons
- Batch Normalization in Practice
- Closing Takeaways on Batch Normalization

2 References

Batch Normalization in Practice

- **Key Point:** Batch normalization has become an essential tool in deep learning due to its ability to stabilize and accelerate training, while also acting as a form of regularization.
- **Impact on Training:** Allows deeper networks to be trained more efficiently and with fewer hyperparameter tuning efforts.

1. Batch Normalization
2. References

[1] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015.

[2] A. Ng and K. Katanforoosh, *CS230 Lecture Notes.*
Stanford University, 2018.

[3] F.-F. Li and Z. Durante, *CS231n Lectures.*
Stanford University, 2024.
Updated June 3, 2024.

Any Questions?