ILLINOIS INSTITUTE OF TECHNOLOGY

Weight standardization analysis

Thomas Ehling A20432671

& Clement Rouault A20432709

CS 577: Deep Learning, IIT Chicago
April 2019

1 Introduction

Deep neural networks have significantly outperform other solution in multiples domains, such as speech, language, vision, games, … One of the reason why deep learning has made such outstanding progress in the recent years is Batch Normalization (BN). BN enables the use of higher learning rates, greatly accelerating the learning process and allowing the use of sigmoid activation functions, previously impossible to train due to the vanishing gradient.

Though batch normalization is now a standard feature of any deep learning framework and can be used off the shelf, using it naively can lead to difficulties in practice. The current recent alternatives to BN are Layer Normalisation (LN), Weight Normalization (WN), Group Normalization (GN), and the latest one : Weight Standardization (WS).

In this paper we briefly explain the bases necessary to understand the project, especially the BN and GN methods. Then, we explain the theory and process behind WS, to finally implement and test it on our own machine learning model.

2 Previous Work

The previous work on the subject are on the regularization techniques related to BN :

- **BN :** Sergey Ioffe, Christian Szegedy (2015) “Batch Normalization: Accelerating Deep Network Training b y Reducing Internal Covariate Shift” : https://arxiv.org/pdf/1502.03167.pdf

- **WN :** Tim Salimans, Diederik P. Kingma (2016) “Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks” : https://arxiv.org/pdf/1602.07868.pdf

- **LN:** Jimmy Lei Ba, Jamie Ryan Kiros, Jamie Ryan Kiros (2016) “Layer Normalization” : https://arxiv.org/pdf/1607.06450.pdf

- **GN:** Yuxin Wu, Kaiming He (2018) “Group Normalization” : https://arxiv.org/pdf/1803.08494.pdf

3 A closer look at the Problem

We aim to understand the WS method, and analyze the performance on real-life applications. The main problem of BN is the low performance of the method on micro-batches, while other solutions does not perform as good for large batches. The searchers claims that when WS is used with GN, it outperform BN everywhere.

The theory part will rely on the research papers for BN, GN and WS.

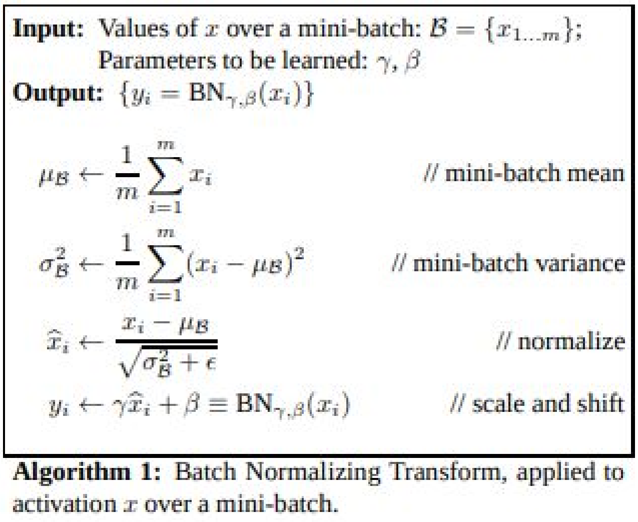
For the experiments, we tested GN+WS vs BN for two different machine learning model :

1. Object detection model on the COCO dataset
2. Classifier on the MNIST Fashion dataset.

4 Theory

4.1 Reminder : BN equations

From the original paper :

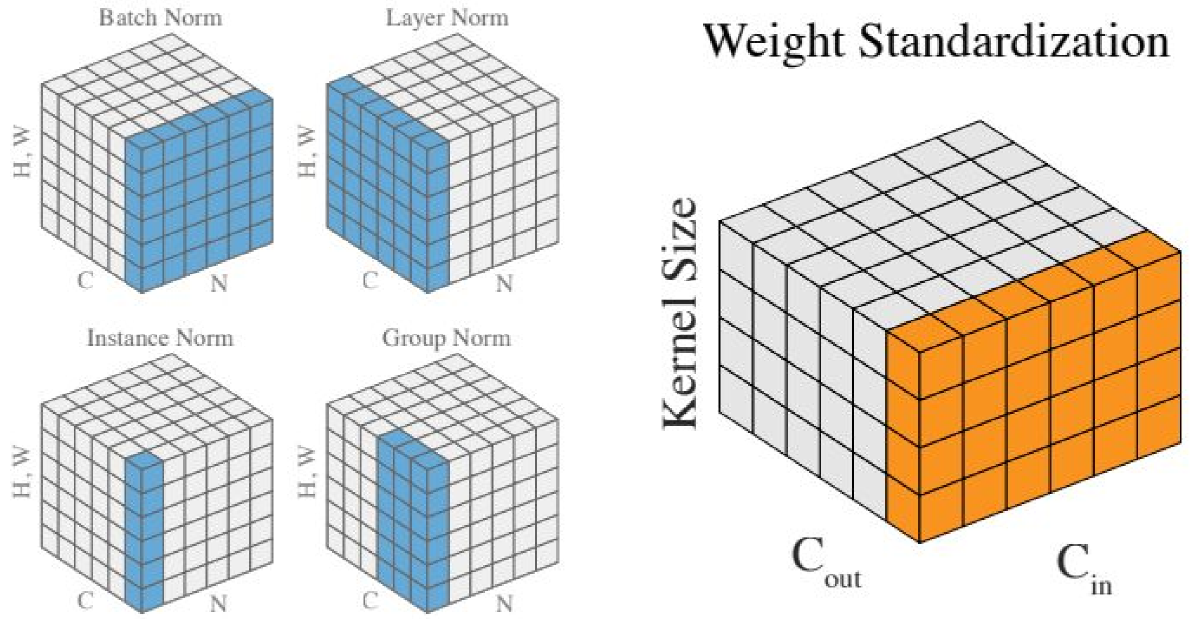


```
Input: Values of $x$ over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
       Parameters to be learned: $\gamma$, $\beta$
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

  $\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$                                // mini-batch mean
  $\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$          // mini-batch variance
  $\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$                   // normalize
  $y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$              // scale and shift
```

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

4.2 Weight Standardization



The figure above illustrate the dimensions impacted by the different regularizations techniques. We can see that the WS technic concern the same dimension as BN, only it affects the weights where BN normalize the batches.

This WS method changes the weights used for the forward path : the new weights used are standardized : they are reduced by the mean, then divided by the standard deviation of the initial weights. It also changes the gradient during backpropagation.

Similar to BN, WS controls the first and second moments of the weights of each output channel individually in convolutional layer.

Important : The searchers assume that the WS is used alongside normalization layers such as GN or BN.

Consider a standard convolutional layer with its bias term set to 0 :

$$\boldsymbol{y} = \hat{\boldsymbol{W}} * \boldsymbol{x}$$

In Weight Standardization, instead of directly optimizing the loss L on the original weights ŵ, the weights ŵ are parameterized as a function of W, i.e., ŵ = WS(W), and optimize the loss L on W by SGD:

$$\hat{\boldsymbol{W}} = \left[\hat{\boldsymbol{W}}_{i,j} \mid \hat{\boldsymbol{W}}_{i,j} = \frac{\boldsymbol{W}_{i,j} - \mu_{\boldsymbol{W}_{i,\cdot}}}{\sigma_{\boldsymbol{W}_{i,\cdot}+\epsilon}}\right] \tag{2}$$

$$\boldsymbol{y} = \hat{\boldsymbol{W}} * \boldsymbol{x} \tag{3}$$

where

$$\mu_{\boldsymbol{W}_{i,\cdot}} = \frac{1}{I}\sum_{j=1}^{I} \boldsymbol{W}_{i,j}, \quad \sigma_{\boldsymbol{W}_{i,\cdot}} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(\boldsymbol{W}_{i,j} - \mu_{\boldsymbol{W}_{i,\cdot}})^2} \tag{4}$$

WS has two main effects :

- **WS normalizes gradients**
- **WS smooth the landscape** by affecting the Lipschitz constant of the Loss.

We are not explaining the details of these processes, these are not a necessity for our project. Please refer to the Research paper in the “Reference” section for more informations.

5 Organization and responsibilities

What was our responsibilities ?

- We **both** work on the **theory** behind Weight Standardization.
- **Thomas** experiment with a **classifier** on the MNIST Fashion dataset
- **Clément** experiment with a **Object detection model** on the COCO dataset

6 Implementation

This part is different for both of us, as we worked on different model.

7 Dataset

This part is different for both of us, as we worked on different model.

8 Model Architecture

This part is different for both of us, as we worked on different model.

9 Evaluation

We merged our individual results for the evaluation part.

Model	Test set	Normalization	Accuracy	Paper Accuracy
	COCO	BN	--%	--%
	COCO	GN+WS	--%	--%
Classifier	MNIST Fashion	BN	--%	X
Classifier	MNIST Fashion	GN+WS	--%	X

- **Concerns and difficulty :**
    - Training time
    - No certainty of efficiency
    - --
    - --
    - --
- **Graph loss and metrics**

10 Conclusion

At first we thought the papers results were too impressive, and we wanted to test if it would perform as well on our own deep learning models. It turns out that our results matches perfectly the ones from the paper. The Weight Standardization coupled with the Group Normalization technique does outperform Batch Normalization.

This is great progress because BN was a breakthrough but it leads searcher to only use large batches, which are memory intensive. This method outperforms BN for micro, standard and large batches, and leads to new opportunities.

We can also notice that all the latest regularization techniques have been developed in the last few years, so we need to watch out for future updates.

References

The original research paper :

- Siyuan Qiao Huiyu Wang Chenxi Liu Wei Shen Alan Yuille : *“Weight Standardization”* (2019)

- https://arxiv.org/pdf/1903.10520v1.pdf

Other references :

- https://github.com/joe-siyuan-qiao/WeightStandardization
- http://mlexplained.com/2018/01/10/an-intuitive-explanation-of-why-batch-normalization-really-works-normalization-in-deep-learning-part-1/
- http://mlexplained.com/2018/01/13/weight-normalization-and-layer-normalization-explained-normalization-in-deep-learning-part-2/
- https://en.wikipedia.org/wiki/Batch_normalization
- https://towardsdatascience.com/batch-normalization-in-neural-networks-1ac91516821c
- https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets