Imperial College London

UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare

Imperial College London

Module: 70152 - Research Tutorial - AI and Machine Learning for Healthcare

Paper Presentation Report

How does Batch Normalisation Help Optimization?

S Santurkar, D Tsipras, A Ilyas and A Madry

Author: Samuel Channon-Wells, M.D.

March 17, 2023

1 Paper Summary

Paper title: How does Batch Normalisation Help Optimization? [1]

Authors: Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry

Publication details: Published at the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.

1.1 Background

Since it was first described by Ioffe and Szegedy in 2015 [2], Batch Normalisation (BN) has demonstrated substantial benefits when training neural networks (NN) [1], [3], and is now a common addition to many NN architectures. These benefits include faster training time [2], [4], reduced learning rates, additional regularisation [5] (often eliminating need for drop-out) and increased robustness to hyperparameter and weight initialisation. The original motivation for introducing BN was to normalise first and second moments of the inputs to activation functions in hidden layers of deep neural networks [2] - so called "Internal Covariate Shift" (ICS) - much akin to how input features are typically normalised across training samples. However, there has been much debate regarding the mechanisms behind the success of BN [6]–[8], and whether or not ICS is responsible.

1.2 Aims of paper

The paper sets out to interrogate the mechanisms responsible for the model improvements seen in Batch Normalisation Neural Networks (BN-NNs). Specifically, the authors explore the relationship between BN-NN improvements and changes in ICS, through a series of empirical experiments. They subsequently present both theoretical and empirical evidence that they claim implies a second mechanism, a "smoothening effect", may be responsible for the improved training efficiency of BN-NNs.

1.3 Methods and results of paper

The first two sections of the paper describe empirical experiments based on two NN architectures. The VGG neural network is a deep convolution neural network that is commonly used for Image classification tasks [9]. For their experiments the authors trained this model on a publicly available image recognition dataset, CIFAR-10 [10]. The second model was simpler, comprising a 25-layer deep linear network, trained on a regression task using an artificially generated Gaussian dataset. The authors rationalise this second choice by conjecturing that it allows them to interrogate the relationship with BN and non-linearities, which are present in the VGG model but not the DLN. After these empirical experiments the authors then spend the final major section of the paper presenting theoretical results, which they claim support their empirical evidence.

The first experiment presented demonstrates how the addition of BN layers into the VGG architecture improves both training and test accuracy, and also facilitates the use of larger learning rates. In parallel, the authors graphically present the distributions of random activations from two layers during the training steps, for both the standard VGG and VGG+BN models. The authors claim this shows that the distributions are highly similar, providing evidence that ICS has not been improved, despite the large improvement in performance. The authors use the above motivating example as a springboard to present further evidence against the ICS explanation for BN. By adding in random noise to perturb the input distributions after the BN layers, the authors demonstrate that even after introducing substantial covariate shift the BN

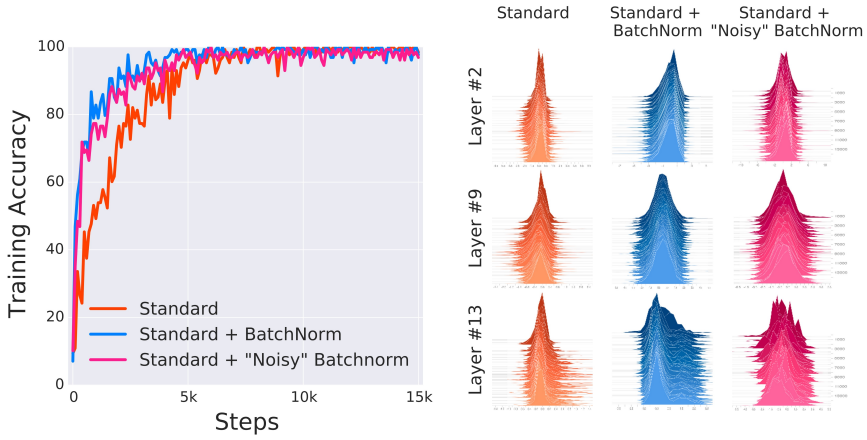
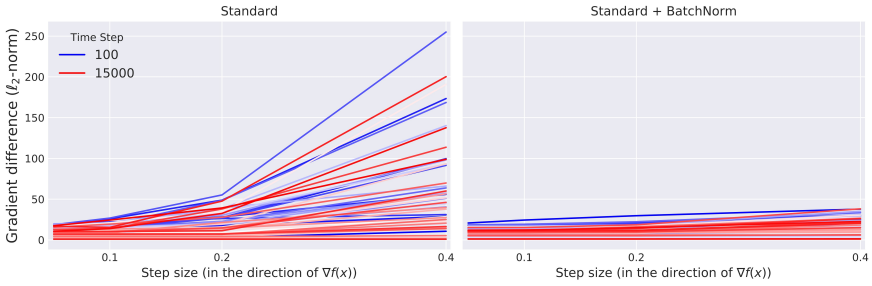


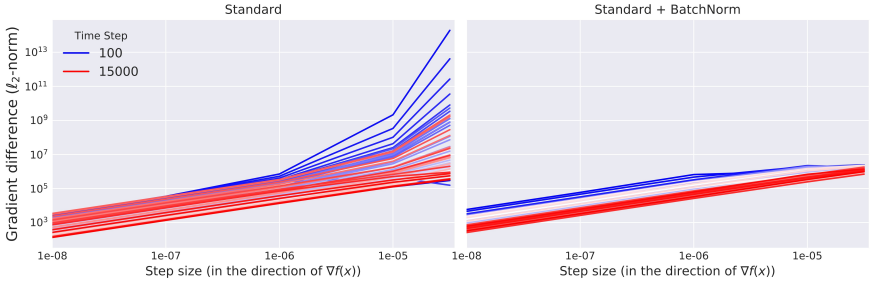
Figure 1: Copy of figure 2 from Santurkar et al. In NeurIPS 2018 [1]

model leads to substantial improvements in training accuracy. Results from this experiment are presented in the 2nd figure, a copy of which is presented below in Figure 1, and 7th figure of the manuscript. Whilst this experiment is compelling, they chose to omit any discussion of whether this also affects learning rates. In addition, they present here only the VGG model, and not the DLN model, without any clear explanation why.

Before moving on to explore the so called "smoothening effect" of BN, the authors conclude the section on ICS by creating their own definition of Internal Covariate Shift based on changes in the loss gradient, motivated by studying the underlying optimisation landscape rather than input distributions. Further empirical evidence is presented in the manuscript's 3rd figure that BN also does not appear to improve the internal covariate shift using their novel definition. Although the reasoning behind this new metric appears sensible, the authors fail at this point to demonstrate empirically that it is an important marker of ICS, or how it relates to the traditional understanding of ICS.



(a) VGG



(b) DLN

Figure 2: Copy of figure 10 from Santurkar et al. In NeurIPS 2018 [1]

Having concluded that ICS is not the underlying mechanism behind BN the second section presents an empirical study of BN's effect on the optimisation landscape. Their first experiment (10th figure from the manuscript, a copy of which is presented below in Figure 2) shows that BN smooths NN loss functions, by reducing the rate at which the gradient changes as we move in the current gradient direction during the gradient-descent step. This effect was seen for early and late time steps, and for both the VGG and DLN models, and I believe gives compelling evidence to support their claims.

By then varying the step sizes taken (equivalent to varying the learning rate) the authors explore the variability in the loss function and its gradient, demonstrating in their 4th figure that introducing BN to the VGG network does indeed lead to a smoother loss function. A copy of this figure is presented in Figure 3. "Smoothness" here is defined in three separate ways: the loss function variability, the reliability of the gradient, and the "effective β-smoothness". BN appears to improve all three metrics of smoothness. However, whilst the results for the VGG model in Figure 3 are convincing, the picture is less clear for the DLN. Results in supplementary figures 9 and 12 of the paper (not presented here) suggest that the smoothness metrics for the DLN, with and without BN, are similar after the first few thousand training steps. In fact, the loss variability and gradient predictiveness appear worse in the BN model. Despite this, the authors still claim that BN smoothening is responsible for the improvements in training.

(a) loss landscape (b) gradient predictiveness (c) "effective" β-smoothness

Figure 3: Copy of figure 4 from Santurkar et al. In NeurIPS 2018 [1]

This section concludes with a final experiment, using different normalisation protocols based on L_p-norms, to show that other normalisation protocols produce similar smoothening effects. Unlike the BN procedure, these methods do not implicitly attempt to control the first two moments of the input distributions. The authors use this to rationalise that the effectiveness of BN may be pure chance, since the motivation for BN (control of these low-order moments) does not appear to be necessary to achieve the observed smoothening properties.

To add theoretical evidence for the smoothening conjecture, the final major section of the paper presents a theoretical analysis by adding a single BN layer to an "arbitrary linear layer" of an otherwise-unspecified Deep Neural Network (DNN). Although the rationale given for this is vaguely described as studying the "reparameterisation of the landscape", I believe such a restrictive oversimplification is likely to have been motivated by mathematical intractability of a system with arbitrarily many BN layers.

In this simplified model, the authors present three main theorems, which essentially provide bounds on the local loss gradient, the local Hessian of the loss, and a "worst-case" local maximum of the gradient function. Under quite broad conditions, these bounds provide evidence that the loss function is indeed smoother after adding the BN layer, with more predictive gradients. Finally, a similar result is presented in Lemma 4.5 (presented in Figure 4) that produces a bound on the distance of the initial network weights from their local minima, suggesting that

BN should improve robustness of weight initialisation through a reduction in this distance.

Lemma 4.5 (BatchNorm leads to a favourable initialization)**.** *Let $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{W}}^*$ be the set of local optima for the weights in the normal and BN networks, respectively. For any initialization W_0*

$$\left|\left|W_0 - \widehat{W}^*\right|\right|^2 \leq ||W_0 - W^*||^2 - \frac{1}{||W^*||^2}\left(||W^*||^2 - \langle W^*, W_0\rangle\right)^2,$$

if $\langle W_0, W^*\rangle > 0$, where $\widehat{W}^*$ and W^* are closest optima for BN and standard network, respectively.

Figure 4: Copy of Lemma 4.5 from Santurkar et al. In NeurIPS 2018 [1]

To provide additional context for their results, the final section comprises a brief summary of related work on normalisation protocols, and other attempts to explain BN's mechanisms. The authors claim that a variety of other normalisation schemes result in similar improvements to BN, although without any analysis to corroborate this claim. They conclude their summary by discussing the paper by Kohler and colleagues [11], which claims yet another explanation for BN's success, namely the disassociation of the length and direction of weights. Whilst this observation is interesting, the authors do not conjecture how this affects their own findings.

After presenting this substantial analysis, the authors conclude that the suggested link between ICS and BN's success is "tenuous at best". Instead, they claim to have identified an important distinct mechanism, by which BN smooths the loss function of the network to allow for more stable and predictable training, avoiding common pitfalls in deep neural network training (e.g. gradient explosion or vanishing). They also claim this smoothing is not unique to BN, but may in fact be a more generic feature of normalisation protocols.

1.4 Critical review

This paper is extremely comprehensive in the depth and breadth of analysis presented, and I believe sheds new light on the phenomenon of Batch Normalisation. In spite of its complexity, I found the paper interesting to read, partly due to the strong question-and-answer narrative used, and simple figures. I particularly enjoyed the mixture of theory and empirical evidence the authors used to substantiate their hypothesis. However, there are some issues regarding the validity of the results which I believe the authors failed to address during the presentation of their work.

My first and principal concern is the theoretical analysis presented, which I believe is limited by oversimplification to a single BN layer. In reality, most networks use BN in multiple layers, and so the authors' theoretical results have unknown validity in this real-world setting. In fact, the work of Yang et al. [12] published the year after this paper (2019) shows that deep networks with multiple BN layers suffer from extensive gradient explosion, highlighting BN as the cause for this issue, in direct contrast to the conclusions of the paper above. One simple explanation for this contradiction is the aforementioned oversimplification. This is supported by Yang et al., who claim that propagation of gradients through multiple BN layers is the main driving factor behind this issue, which can be alleviated through adding skip-connections.

In general, the figures were simple, intuitive, and provided inciteful explanations. The main figures provide convincing evidence that, for the VGG model in particular, BN does indeed improve model performance and training, without addressing ICS, and supports the key hypothesis of the paper; that smoothening of the loss function is an important mechanism underlying BN's success.

However, the authors choose to present figures for the VGG model more prominently in the

paper, relegating comparative figures for the DLN to the supplement. I believe this warrants explanation, especially as the results for the DLN model often appeared less supportive of the authors' overall conclusions. This introduces speculation regarding reporting bias. Similarly, the authors provided no clear explanation for their choices of the two models used, and did not compare the model results in great detail, or comment on how this choice affects the generalisability of their results. Without testing in additional NN architectures the conclusions of the paper are therefore likely overstated, as the results could be due to specific characteristics of the limited models presented.

A great strength of the paper was its strong narrative, improving the readability despite in-depth methodological descriptions. Unfortunately, the description in the section on additional normalisation protocols was vague, and limits the replicability of the work. In addition, several of the figure captions were not descriptive enough, and it was therefore not always clear how figures should be interpreted by the reader.

In spite of these limitations, I found the paper inciteful and thought-provoking. I believe the authors provide a compelling argument that, in certain contexts, a key feature of BN's success is its loss-smoothing properties, but I am sceptical about the generalisability of these findings. I believe the true explanation for BNs success are likely multifactorial, and may depend on the underlying architecture it is applied to.

2 Discussion and Reflection

2.1 Tutorial discussion

As the paper was heavily theoretical the focus of the peer discussion was on clarification of technical aspects of the paper. This was despite a detailed overview of Batch Normalisation given at the start of my presentation. For example, one colleague asked how the batch size might affect firstly general BN performance, and secondly the results of the review paper. From my background reading I was able to explain the problem of overfitting seen in BN networks trained with small batch-sizes [4]. However, I was not able to provide an answer to the second question, which was also not addressed in the paper. We also discussed other limitations of the paper, including the lack of clarity in certain figure captions, and how this can really decrease the readability and impact of published work.

For me the most interesting discussion point was regarding the new definition of Internal Covariate Shift (ICS) introduced in the 1st section of the paper. During my presentation I remarked that the lack of external use of this definition may limit its usefulness to quantify ICS - which was originally defined very differently. The subsequent discussion led me to question if the conclusions are therefore misleading, but also to reflect on the challenges of presenting research in an area where nomenclature is not universally defined. In such situations I believe it is important to be aware of the limitations of using non-standard nomenclature, and to report these clearly. Notably, I feel the paper in question did not discuss this in enough detail.

One specific question regarding the BN procedure addressed the issue of using BN layers before or after activation functions. We agreed as a group that this was likely to affect the optimisation problem, and hence model performance, but we were unable as a group to determine how this might affect the results of the paper. On reflection, this highlights the difficult issue of balancing concise presentation of relevant results with ensuring enough results are presented to address multiple possible model configurations. I feel the reviewed paper achieved this balance reasonably well, although I would like to have seen more detailed descriptions of figures and methods in the supplements, as well as additional reasoning behind modelling decisions.

2.2 Personal reflections

I enjoyed this challenge immensely, and it was a fantastic opportunity to further develop the critical-thinking approach to published research that we have been developing this term. I chose a difficult paper on an unfamiliar topic in order to challenge myself. As a result, I learnt a huge amount about both the topic itself, and the style and content of Artificial Intelligence/Machine Learning publications. I am relatively new to reading these more in-depth AI publications, which put me outside of my comfort zone. Because of this, I devoted a sizeable proportion of my presentation to providing an overview of Batch Normalisation as I understand it. I was pleased that this was so well received, and this was commented on both in the formal feedback and after the tutorial.

I was also pleased with the feedback. General comments were positive, highlighting my good pacing of the presentation and well-structured narrative. Areas for improvement were focussed on giving a broader overview of other normalisation methods and updates since this paper was published, as well as providing more detail when explaining the complex mathematical aspects of the paper. However, these comments were contrasted with the views of other colleagues, who felt the detail was at times too in-depth. I did struggle with this issue whilst preparing my presentation, but I was pleased with the balance I achieved overall, given the complexity and density of this paper.

3 Funding

This work was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. EP/S023283/1) and European Union's Horizon 2020 program (under GA number 848196 DIAMONDS https://www.diamonds2020.eu/).

References

[1] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf.

[2] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. arXiv: 1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167.

[3] J. Bjorck, C. P. Gomes, and B. Selman, "Understanding batch normalization," *CoRR*, vol. abs/1806.02375, 2018. arXiv: 1806.02375. [Online]. Available: http://arxiv.org/abs/1806.02375.

[4] X. Lian and J. Liu, "Revisit batch normalization: New understanding and refinement via composition optimization," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, K. Chaudhuri and M. Sugiyama, Eds., ser. Proceedings of Machine Learning Research, vol. 89, PMLR, 16–18 Apr 2019, pp. 3254–3263. [Online]. Available: https://proceedings.mlr.press/v89/lian19a.html.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *CoRR*, vol. abs/1611.03530, 2016. arXiv: 1611.03530. [Online]. Available: http://arxiv.org/abs/1611.03530.

[6] A. S. Morcos, D. G. T. Barrett, N. C. Rabinowitz, and M. Botvinick, *On the importance of single directions for generalization*, 2018. DOI: 10.48550/ARXIV.1803.06959. [Online]. Available: https://arxiv.org/abs/1803.06959.

[7] A. Rahimi and B. Recht, "Back when we were kids," *NIPS Test of Time Award*, 2017.

[8] *Batch norm explained visually - why does it work?* Web Page, 2021. [Online]. Available: https://towardsdatascience.com/batch-norm-explained-visually-why-does-it-work-90b98bcc58a0.

[9] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014. DOI: 10.48550/ARXIV.1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556.

[10] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[11] J. Kohler, H. Daneshmand, A. Lucchi, M. Zhou, K. Neymeyr, and T. Hofmann, *Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization*, 2018. DOI: 10.48550/ARXIV.1805.10694. [Online]. Available: https://arxiv.org/abs/1805.10694.

[12] G. Yang, J. Pennington, V. Rao, J. Sohl-Dickstein, and S. S. Schoenholz, "A mean field theory of batch normalization," *CoRR*, vol. abs/1902.08129, 2019. arXiv: 1902.08129. [Online]. Available: http://arxiv.org/abs/1902.08129.