Normalization

Sarthak Patel
School of Physical Sciences, National Institute of Science Education and Research, Bhubaneswar, India
Integrated MSc. , 5th Year, 1911149

(CS460 – Machine Learning)

Instructor - Dr.Shubhankar Mishra

Introduction

Neural networks may struggle when working with input features that have different value ranges, causing longer training times and convergence issues. Training models on unnormalized data leads to slower convergence due to gradient descent difficulties. The problem of exploding gradients arises, causing instability as high values propagate through network layers. Preprocessing techniques like normalization and standardization help overcome scaling issues.

Normalization

- Normalization is a data preprocessing technique that transforms input feature values to the range [0,1].
- The transformation for a particular input feature x, that has values in the range [x_{min} , x_{max}] is represented using the equation,

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization

- Standardization is a data preprocessing technique that transforms input values to follow a distribution with a zero mean (μ) and unit variance (σ^2).
- It achieves this by using a mathematical formula ,

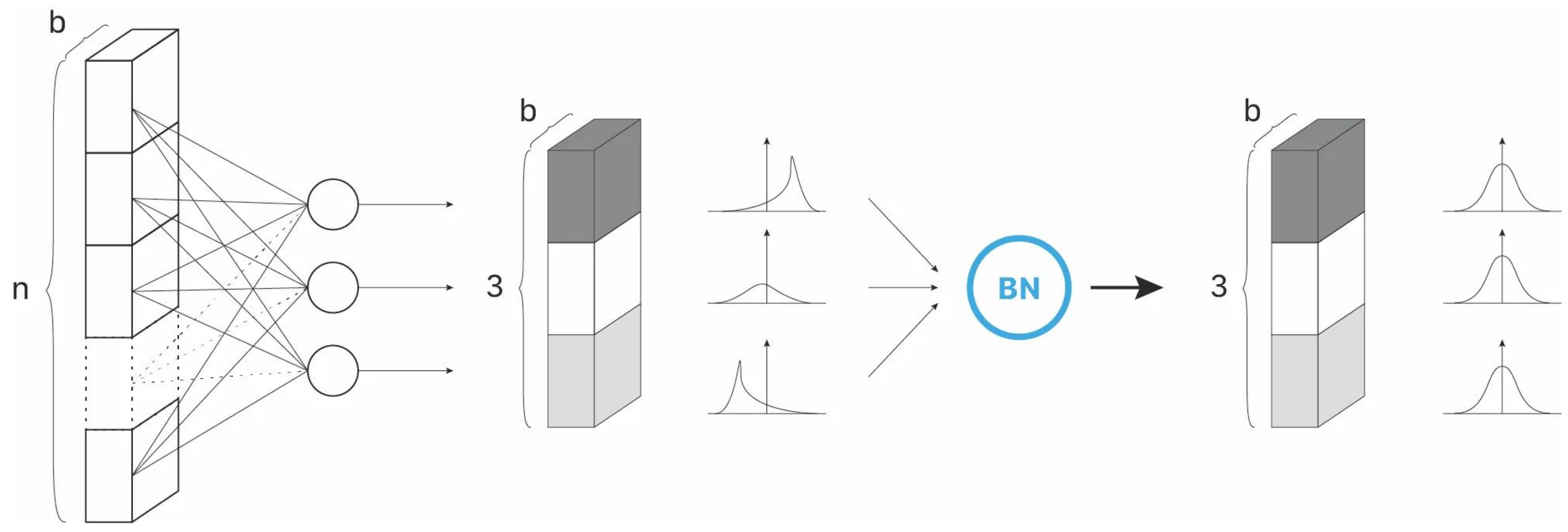
$$x_{std} = \frac{x - \mu}{\sigma}$$

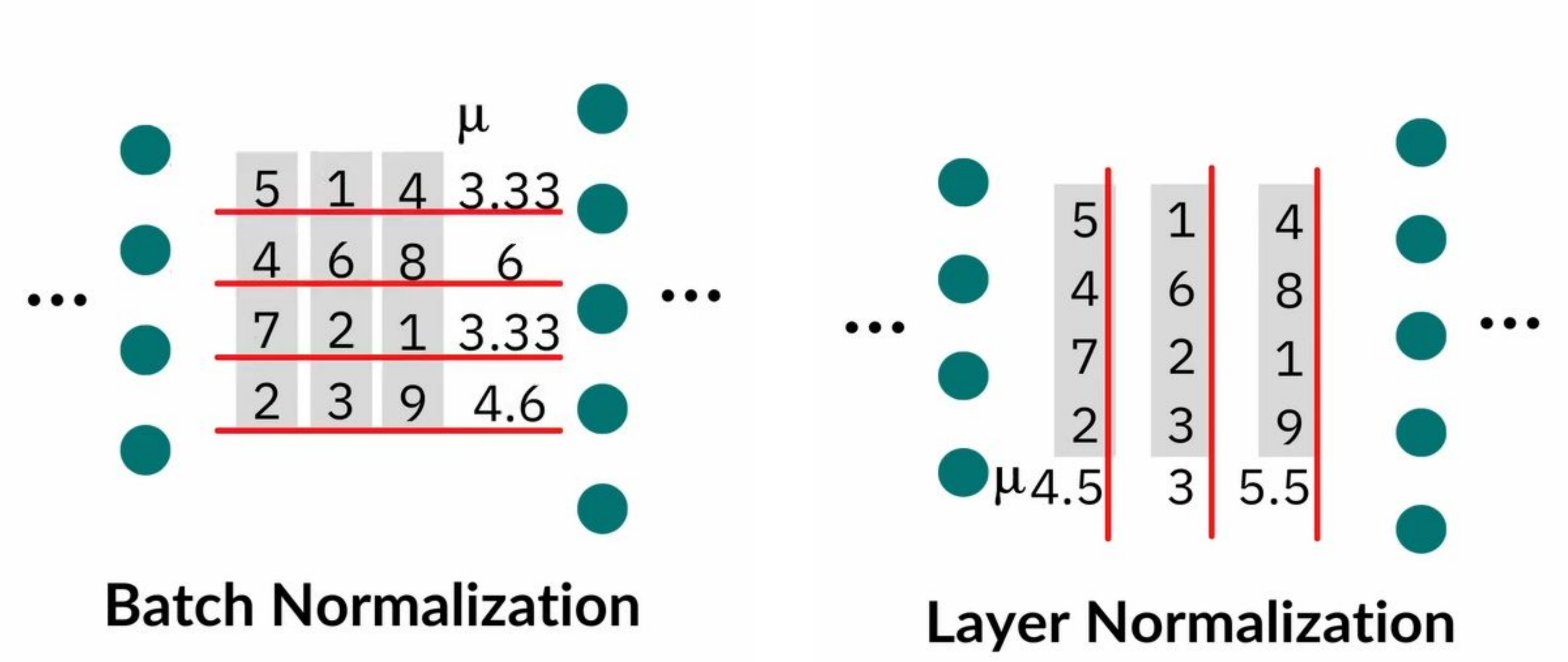
Batch Normalization

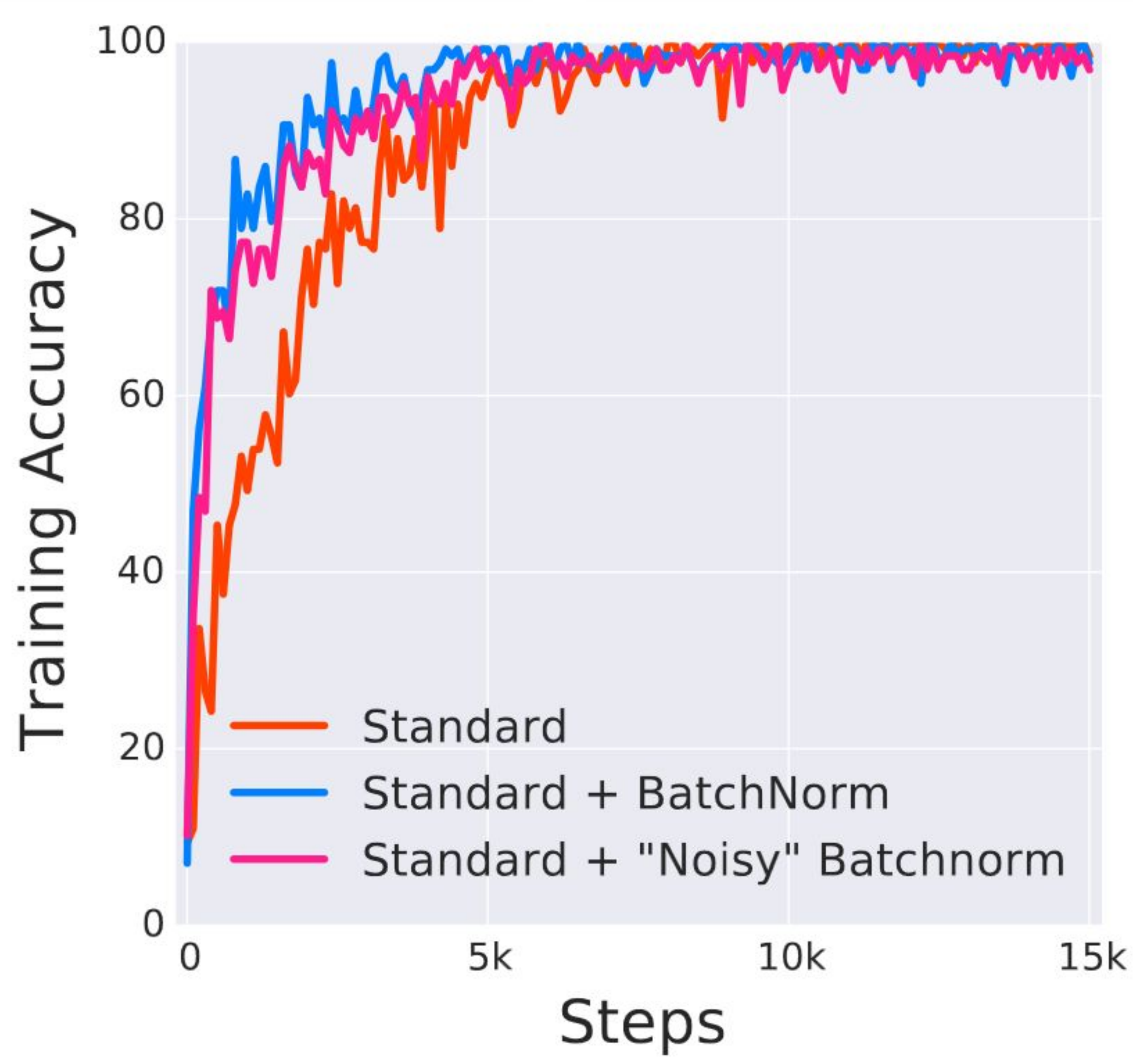
- Batch normalization (BN) involves normalizing activation vectors in hidden layers using the mean and variance of the current batch's data.
- This normalization is implemented right before or after the application of the nonlinear function.
- Batch normalization is less suitable for sequence models and smaller batch sizes.

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

$$\overline{x_i} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \qquad y_i = \gamma\overline{x_i} + \beta$$







Saturkar et al. (2019) demonstrate in a VGG network using the CIFAR10 dataset that training accuracy is improved when employing Batch Normalization

Layer Normalization

- In Layer Normalization, all neurons within a layer share the same distribution across all input features.
- Layer normalization is independent of the batch size, so it can be applied to batches with smaller sizes as well.

$$\mu_l = \frac{1}{d}\sum_{i=1}^{d} x_i (1) \qquad \sigma_l^2 = \frac{1}{d}\sum_{i=1}^{d}(x_i - \mu_l)^2 \,(2)$$

$$\hat{x}_i = \frac{x_i - \mu_l}{\sqrt{\sigma_l^2}} (3) \qquad y_i = \mathcal{LN}\,(x_i) = \gamma.\hat{x}_i + \beta (4)$$

References

- Santurkar, S., Tsipras, D., Ilyas, A., & Madry, A. (2019). How Does Batch Normalization Help Optimization?. stat, 1050, 15.
- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.
- https://www.assemblyai.com/blog/what-is-layer-normalization/
- https://towardsdatascience.com/batch-normalization-in-3-levels-of-understanding-14c2da90a338

Created with Libre Office Impress and Google slides