3-3 Batch Normalization

Zhonglei Wang

WISE and SOE, XMU, 2025

Contents

Introduction

1. Proposed by Ioffe and Szegedy (2015)
   - Tries to solves an interval covariate shift problem
   - Normalize values after linear transformation but before activation for each neuron
   - Introduce two more parameters to allow for heterogeneity

Forward propagation

1. For a mini-batch with m training examples, the forward propagation for the lth layer is

$$\boldsymbol{Z}^{[l]} = \boldsymbol{A}^{[l-1]}(\boldsymbol{W}^{[l]})^{\mathrm{T}} + (\boldsymbol{b}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m\times d^{[l]}}; \quad \boldsymbol{A}^{[l]} = \sigma^{[l]}(\boldsymbol{Z}^{[l]}) \in \mathbb{R}^{m\times d^{[l]}}$$

2. Denote $\boldsymbol{Z}^{[l]} = (\boldsymbol{z}_1^{[l]}, \ldots, \boldsymbol{z}_m^{[l]})^{\mathrm{T}}$

3. After linear transformation, we consider the following new calculations:

$$\boldsymbol{\mu}^{[l]} = m^{-1}(\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]}\times 1}$$

$$\boldsymbol{\sigma}^{2[l]} = m^{-1}\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right) \circ \left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right) \in \mathbb{R}^{d^{[l]}\times 1}$$

$$\tilde{\boldsymbol{z}}_{i,norm}^{[l]} = \frac{\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}}{\sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon}} \in \mathbb{R}^{d^{[l]}\times 1}$$

$$\tilde{\boldsymbol{z}}_i^{[l]} = \boldsymbol{\gamma}^{[l]} \circ \tilde{\boldsymbol{z}}_{i,norm}^{[l]} + \boldsymbol{\beta}^{[l]} \in \mathbb{R}^{d^{[l]}\times 1}$$

$$\tilde{\boldsymbol{Z}}^{[l]} = (\tilde{\boldsymbol{z}}_1^{[l]}, \ldots, \tilde{\boldsymbol{z}}_m^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m\times d^{[l]}}$$

Forward propagation

1. Vectorization for the red calculations, but leave the blue parts alone

$$\boldsymbol{Z}^{[l]} = \boldsymbol{A}^{[l-1]}(\boldsymbol{W}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m \times d^{[l]}}$$

$$\boldsymbol{\mu}^{[l]} = m^{-1}(\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]} \times 1}$$

$$\check{\boldsymbol{Z}}^{[l]} = \boldsymbol{Z}^{[l]} - \mathbf{1}(\boldsymbol{\mu}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m \times d^{[l]}}$$

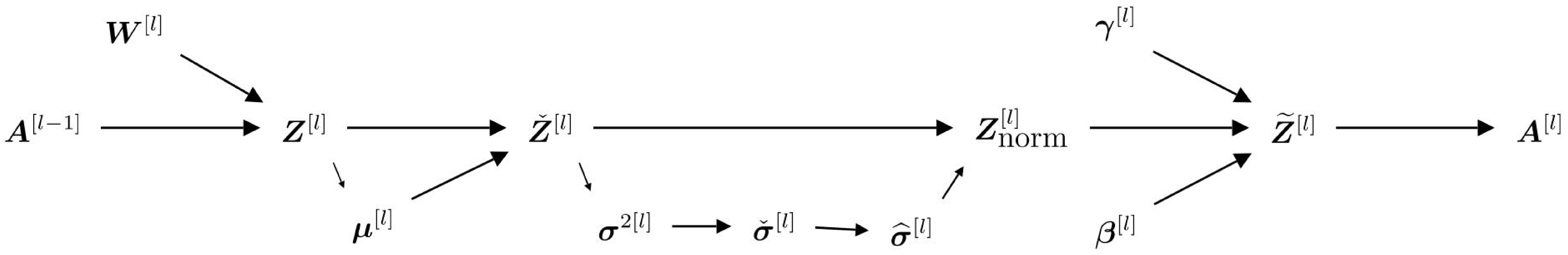
$$\boldsymbol{\sigma}^{2[l]} = m^{-1}\sum_{i=1}^{m} \check{\boldsymbol{z}}_i^{[l]} \circ \check{\boldsymbol{z}}_i^{[l]} \in \mathbb{R}^{d^{[l]} \times 1}$$

$$\check{\boldsymbol{\sigma}}^{[l]} = \sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon} \in \mathbb{R}^{d^{[l]} \times 1}$$

$$\widehat{\boldsymbol{\sigma}}^{[l]} = (\check{\boldsymbol{\sigma}}^{[l]})^{-1} \in \mathbb{R}^{d^{[l]} \times 1}$$

$$\boldsymbol{Z}_{\text{norm}}^{[l]} = \check{\boldsymbol{Z}}^{[l]} \circ \{\mathbf{1}(\widehat{\boldsymbol{\sigma}}^{[l]})^{\mathrm{T}}\} \in \mathbb{R}^{m \times d^{[l]}}$$

2. Notice that the previous bias term $\boldsymbol{b}^{[l]}$ is useless for batch normalization



Forward propagation for the red parts:

$$\boldsymbol{Z}^{[l]} = \boldsymbol{A}^{[l-1]}(\boldsymbol{W}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m\times d^{[l]}}$$

$$\boldsymbol{\mu}^{[l]} = (\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]}\times 1}$$

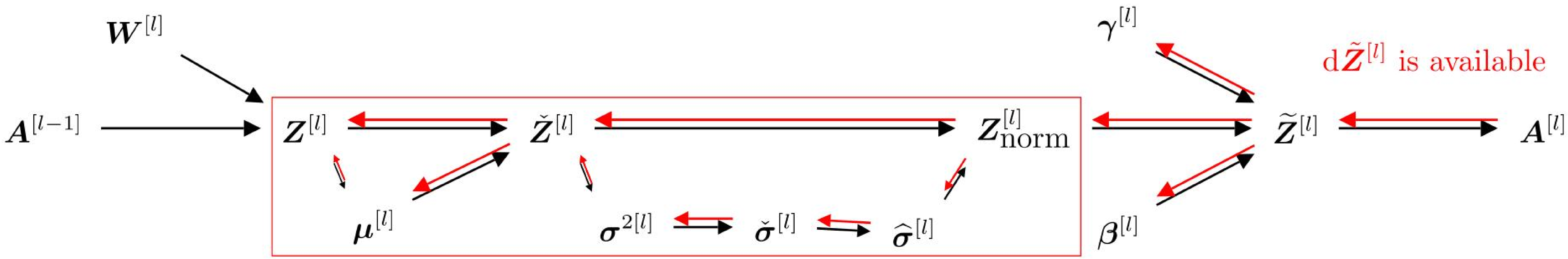
$$\check{\boldsymbol{Z}}^{[l]} = \boldsymbol{Z}^{[l]} - \mathbf{1}(\boldsymbol{\mu}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m\times d^{[l]}}$$

$$\boldsymbol{\sigma}^{2[l]} = m^{-1}\sum_{i=1}^{m} \check{\boldsymbol{z}}_i^{[l]} \circ \check{\boldsymbol{z}}_i^{[l]} \in \mathbb{R}^{d^{[l]}\times 1}$$

$$\check{\boldsymbol{\sigma}}^{[l]} = \sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon} \in \mathbb{R}^{d^{[l]}\times 1}$$

$$\widehat{\boldsymbol{\sigma}}^{[l]} = (\check{\boldsymbol{\sigma}}^{[l]})^{-1} \in \mathbb{R}^{d^{[l]}\times 1}$$

$$\boldsymbol{Z}_{\text{norm}}^{[l]} = \check{\boldsymbol{Z}}^{[l]} \circ \{\mathbf{1}(\widehat{\boldsymbol{\sigma}}^{[l]})^{\mathrm{T}}\} \in \mathbb{R}^{m\times d^{[l]}}$$



Backpropagation:

$$\mathrm{d}\boldsymbol{Z}^{[l]} = \mathrm{d}\boldsymbol{Z}_1^{[l]} + \mathrm{d}\boldsymbol{Z}_2^{[l]} \qquad \mathrm{d}\boldsymbol{\beta}^{[l]} = \left(\mathrm{d}\tilde{\boldsymbol{Z}}^{[l]}\right)^{\mathrm{T}} \mathbf{1} \qquad \mathrm{d}\boldsymbol{\gamma}^{[l]} = \left(\mathrm{d}\tilde{\boldsymbol{Z}}^{[l]} \circ \boldsymbol{Z}_{\text{norm}}^{[l]}\right)^{\mathrm{T}} \mathbf{1}$$

$$\mathrm{d}\check{\boldsymbol{Z}}^{[l]} = \mathrm{d}\check{\boldsymbol{Z}}_1^{[l]} + \mathrm{d}\check{\boldsymbol{Z}}_2^{[l]}$$

$$\mathrm{d}\boldsymbol{Z}_1^{[l]} = \mathrm{d}\check{\boldsymbol{Z}}^{[l]}$$

$$\mathrm{d}\boldsymbol{\mu}^{[l]} = -\left(\mathrm{d}\check{\boldsymbol{Z}}^{[l]}\right)^{\mathrm{T}} \mathbf{1}$$

$$\mathrm{d}\boldsymbol{Z}_2^{[l]} = m^{-1} \mathbf{1} \left(\mathrm{d}\boldsymbol{\mu}^{[l]}\right)^{\mathrm{T}}$$

$$\mathrm{d}\check{\boldsymbol{\sigma}}^{[l]} = -\mathrm{d}\widehat{\boldsymbol{\sigma}}^{[l]} \circ \widehat{\boldsymbol{\sigma}}^{[l]} \circ \widehat{\boldsymbol{\sigma}}^{[l]}$$

$$\mathrm{d}\boldsymbol{\sigma}^{2[l]} = \mathrm{d}\check{\boldsymbol{\sigma}}^{[l]} \circ \widehat{\boldsymbol{\sigma}}^{[l]}/2$$

$$\mathrm{d}\check{\boldsymbol{Z}}_2^{[l]} = 2m^{-1} \check{\boldsymbol{Z}}^{[l]} \circ \left\{\mathbf{1}\left(\mathrm{d}\boldsymbol{\sigma}^{2[l]}\right)^{\mathrm{T}}\right\}$$

$$\mathrm{d}\boldsymbol{Z}_{\text{norm}}^{[l]} = \mathrm{d}\tilde{\boldsymbol{Z}}^{[l]} \circ \left\{\mathbf{1}\left(\boldsymbol{\gamma}^{[l]}\right)^{\mathrm{T}}\right\}$$

$$\mathrm{d}\check{\boldsymbol{Z}}_1^{[l]} = \mathrm{d}\boldsymbol{Z}_{\text{norm}}^{[l]} \circ \left\{\mathbf{1}\left(\widehat{\boldsymbol{\sigma}}^{[l]}\right)^{\mathrm{T}}\right\}$$

$$\mathrm{d}\widehat{\boldsymbol{\sigma}}^{[l]} = \left(\mathrm{d}\boldsymbol{Z}_{\text{norm}}^{[l]} \circ \check{\boldsymbol{Z}}^{[l]}\right)^{\mathrm{T}} \mathbf{1}$$

Remarks

1. Disadvantage
    - Since we normalize "inputs" before activation, many different "inputs" may result in same "activations"
    - Besides, batch normalization also introduces more model parameters

Remarks

1. Advantage

   - Stablize forward propogation
   - Stablize forward propogation
     - ▷ The variance can be controlled by γ's
   - Higher learning rates
     - ▷ Batch normalization can make loss and its gradients more smooth
   - Regularization
     - ▷ We injects noises from other training examples throught mean and variance
     - ▷ Thus, batch normalization may improve the generality of the network