## 3-3 Batch Normalization

Zhonglei Wang WISE and SOE, XMU, 2025

## Contents

1. Introduction

2. Forward propagation

3. Backpropagation

# 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

Wang, Z. (WISE & SOE, XMU)

3-3 Batch Normalization

# Forward propagation

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

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

- 2. Denote  ${m Z}^{[l]} = ({m z}_1^{[l]}, \dots, {m z}_m^{[l]})^{\mathrm{T}}$
- 3. After linear transformation, we consider the following new calculations:

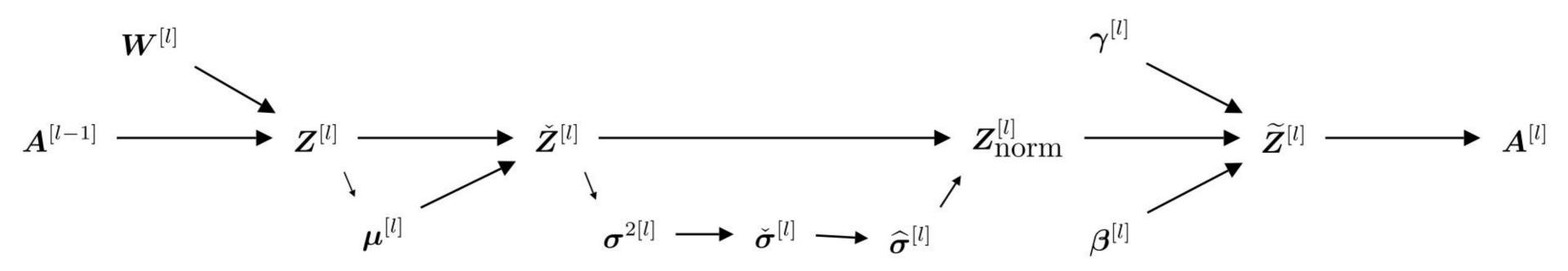
$$\begin{split} \boldsymbol{\mu}^{[l]} &= m^{-1}(\boldsymbol{Z}^{[l]})^{\mathrm{T}} \mathbf{1} \in \mathbb{R}^{d^{[l]} \times 1} \\ \boldsymbol{\sigma}^{2[l]} &= m^{-1} \sum_{i=1}^{m} \left( \boldsymbol{z}_{i}^{[l]} - \boldsymbol{\mu}^{[l]} \right) \circ \left( \boldsymbol{z}_{i}^{[l]} - \boldsymbol{\mu}^{[l]} \right) \in \mathbb{R}^{d^{[l]} \times 1} \\ \tilde{\boldsymbol{z}}_{i,norm}^{[l]} &= \frac{\boldsymbol{z}_{i}^{[l]} - \boldsymbol{\mu}^{[l]}}{\sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon}} \in \mathbb{R}^{d^{[l]} \times 1} \\ \tilde{\boldsymbol{z}}_{i}^{[l]} &= \boldsymbol{\gamma}^{[l]} \circ \tilde{\boldsymbol{z}}_{i,norm}^{[l]} + \boldsymbol{\beta}^{[l]} \in \mathbb{R}^{d^{[l]} \times 1} \\ \tilde{\boldsymbol{z}}^{[l]} &= (\tilde{\boldsymbol{z}}_{1}^{[l]}, \dots, \tilde{\boldsymbol{z}}_{m}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m \times d^{[l]}} \end{split}$$

# Forward propagation

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

$$egin{aligned} oldsymbol{Z}^{[l]} &= oldsymbol{A}^{[l-1]}(oldsymbol{W}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m imes d^{[l]}} \ oldsymbol{\mu}^{[l]} &= m^{-1}(oldsymbol{Z}^{[l]})^{\mathrm{T}} oldsymbol{1} \in \mathbb{R}^{d^{[l]} imes 1} \ oldsymbol{\check{Z}}^{[l]} &= oldsymbol{Z}^{[l]} - oldsymbol{1}(oldsymbol{\mu}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m imes d^{[l]}} \ oldsymbol{\sigma}^{2[l]} &= m^{-1} \sum_{i=1}^{m} oldsymbol{\check{z}}^{[l]}_i \circ oldsymbol{\check{z}}^{[l]}_i \in \mathbb{R}^{d^{[l]} imes 1} \ oldsymbol{\check{\sigma}}^{[l]} &= \sqrt{oldsymbol{\sigma}^{2[l]} + \epsilon} \in \mathbb{R}^{d^{[l]} imes 1} \ oldsymbol{\check{\sigma}}^{[l]} &= (oldsymbol{\check{\sigma}}^{[l]})^{-1} \in \mathbb{R}^{d^{[l]} imes 1} \ oldsymbol{Z}^{[l]}_{\mathrm{norm}} &= oldsymbol{\check{Z}}^{[l]} \circ \{ oldsymbol{1}(\widehat{oldsymbol{\sigma}}^{[l]})^{\mathrm{T}} \} \in \mathbb{R}^{m imes d^{[l]}} \end{aligned}$$

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



### Forward propagation for the red parts:

$$oldsymbol{Z}^{[l]} = oldsymbol{A}^{[l-1]} (oldsymbol{W}^{[l]})^{ ext{T}} \in \mathbb{R}^{m imes d^{[l]}}$$

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

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

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

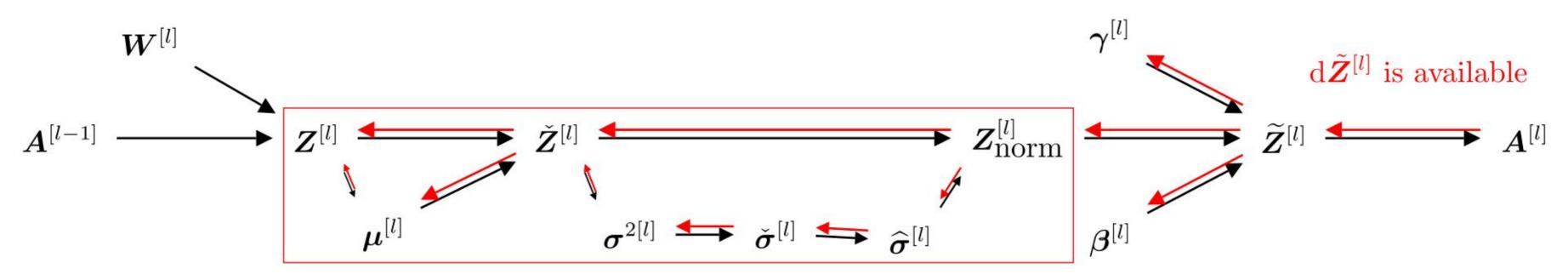
$$\check{\boldsymbol{\sigma}}^{[l]} = \sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon} \in \mathbb{R}^{d^{[l]} \times 1}$$

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

$$m{Z}_{ ext{norm}}^{[l]} = \check{m{Z}}^{[l]} \circ \{ m{1} (\widehat{m{\sigma}}^{[l]})^{ ext{T}} \} \in \mathbb{R}^{m imes d^{[l]}}$$

Wang, Z. (WISE & SOE, XMU)=1

3-3 Batch Normalization



### Backpropagation:

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

$$\mathrm{d}oldsymbol{eta}^{[l]} = \left(\mathrm{d} ilde{oldsymbol{Z}}^{[l]}
ight)^{\mathrm{T}}\mathbf{1}$$

$$\mathrm{d}oldsymbol{\gamma}^{[l]} = \left(\mathrm{d} ilde{oldsymbol{Z}}^{[l]} \circ oldsymbol{Z}_{\mathrm{norm}}^{[l]}
ight)^{\mathrm{T}} \mathbf{1}$$

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

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

$$\mathrm{d}oldsymbol{Z}_{\mathrm{norm}}^{[l]} = \mathrm{d} ilde{oldsymbol{Z}}^{[l]} \circ \left\{ \mathbf{1} \left(oldsymbol{\gamma}^{[l]}
ight)^{\mathrm{T}} 
ight\}$$

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

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

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

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

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

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

$$_{ ext{Wang, Z. (WISE \& SO}} \mathbf{z}_{\mathbf{z}_{\mathbf{z}_{\mathbf{z}_{\mathbf{z}}}}}^{[l]} = m^{-1} \mathbf{1} \left( \mathrm{d} \boldsymbol{\mu}^{[l]} \right)^{\mathrm{T}}$$

## 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

Wang, Z. (WISE & SOE, XMU)

3-3 Batch Normalization

## Remarks

#### 1. Advantage

- Stablize forward propogation
- Stablize forward propogation
  - ightharpoonup The variance can be controlled by  $\gamma$ '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