3-3 批量归一化

王中雷

厦门大学王亚南经济研究院和经济学院, 2025

内容摘要

1. 简介

2. 前向传播

3. 后向传播

简介

1. 被 Ioffe and Szegedy (2015) 提出

- 可用于解决协变量偏差的问题
- 在线性变换后但在激活运算前进行归一化
- 基于两个额外模型参数，调整特征之间的异质性

前向传播

1. 对于小批量梯度下降法（批量大小为 m），第 l 层的计算为：

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

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

3. 线性变换后，我们考虑下面的新运算：

$$\begin{aligned}
\boldsymbol{\mu}^{[l]} &= m^{-1}(\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]}\times 1} \\
\boldsymbol{\sigma}^{2[l]} &= m^{-1}\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right)\circ\left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right) \in \mathbb{R}^{d^{[l]}\times 1} \\
\tilde{\boldsymbol{z}}_{i,norm}^{[l]} &= \frac{\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}}{\sqrt{\boldsymbol{\sigma}^{2[l]} + \epsilon}} \in \mathbb{R}^{d^{[l]}\times 1} \\
\tilde{\boldsymbol{z}}_i^{[l]} &= \boldsymbol{\gamma}^{[l]} \circ \tilde{\boldsymbol{z}}_{i,norm}^{[l]} + \boldsymbol{\beta}^{[l]} \in \mathbb{R}^{d^{[l]}\times 1} \\
\tilde{\boldsymbol{Z}}^{[l]} &= (\tilde{\boldsymbol{z}}_1^{[l]}, \ldots, \tilde{\boldsymbol{z}}_m^{[l]})^{\mathrm{T}} \in \mathbb{R}^{m\times d^{[l]}}
\end{aligned}$$

前向传播

1．考虑红色计算的向量化，暂时不考虑蓝色部分的计算细节

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

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

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

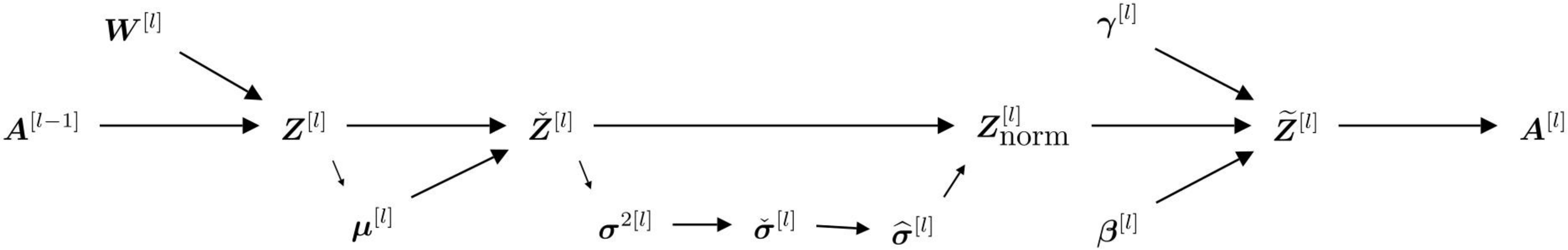
$$\boldsymbol{\sigma}^{2[l]} = m^{-1}\sum_{i=1}^{m} \check{\boldsymbol{z}}_i^{[l]} \circ \check{\boldsymbol{z}}_i^{[l]} \in \mathbb{R}^{d^{[l]}\times 1}$$

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

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

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

2．注意：偏置项 $\boldsymbol{b}^{[l]}$ 对于批量归一化是无用的



红色计算部分的前向传播过程为：

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

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

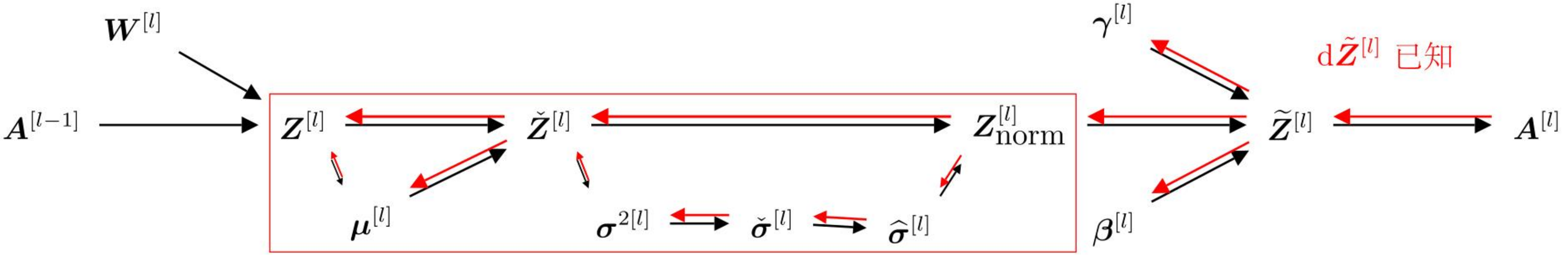
$$\boldsymbol{\mu}^{[l]} = (\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]} \times 1}$$

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

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

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

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



后向传播：

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

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

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

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

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

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

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

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

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

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

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

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

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

注意

1．缺点

- 由于我们在激活前使用批量归一化，不同的“输入特征”可能有相同的“激活值”
- 此外，批量归一化对应着更多的模型参数
- 为什么不在激活运算之后进行批量归一化?

注意

1．优点：

- 稳定化前向传播
  - 批量归一化的方差被γ 控制
- 可以接受更大的学习率
  - 批量归一化可使代价函数及其梯度可以更平滑
- 正则化
  - 批量归一化可被认为将其他训练样本的随机性加入到每个训练样本
  - 因此，批量归一化可以提升神经网络模型的泛化性