人工智能与深度学习

3-3 批量归一化

王中雷

经济学院、王亚南经济研究院、邹至庄经济研究院

厦门大学

（第一版）

内容摘要

1. 简介

2. 前向传播

3. 后向传播

4. 关于 dropout 和批量归一化的讨论

简介

1. 2015 年以前，深度学习模型的训练存在严重的优化困难
   - 后向传播的不稳定
   - 网络对参数初始值极端敏感
2. 例如，一个常被讨论的问题是内部协变量漂移（internal covariate shift）
   - 较浅层模型参数的微小更新，可能会对深层输入的分布，产生较大的影响
   - 较深层的模型参数，始终在“追逐移动目标”
3. 这种不稳定，主要表现为下面的两个问题
   - 梯度消失：当线性变换结果，漂移到激活函数的饱和区时，局部梯度趋近于 0
   - 梯度爆炸：权重初始化过大，则激活的方差会随层数放大，导致代价函数发散

简介

1. 对训练集进行“白化（whitening）”，是经典统计学习模型的常规操作
   - 将输入数据变为零均值、单位方差，且特征间互不相关
   - 算法收敛可能会更快
2. 上面的白化操作，对于深度学习而言不可行
   - 训练集的规模往往较大
   - 我们不能确保每一层都能实现白化操作
   - 白化操作的计算成本高

简介

1. Ioffe and Szegedy (2015) 提出了批量归一化 (Batch normalization)
   - 逐特征独立处理：对每个（神经元输出的）特征独立进行归一化，而不考虑特征之间的协方差
   - 小批量统计估计：使用当前 mini-batch，得到所需的均值和方差
2. 在线性变换后，但在激活运算前，实施批量归一化
3. 批量归一化可实现如下效果
   - 可用于解决内部协变量漂移的问题
   - 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：

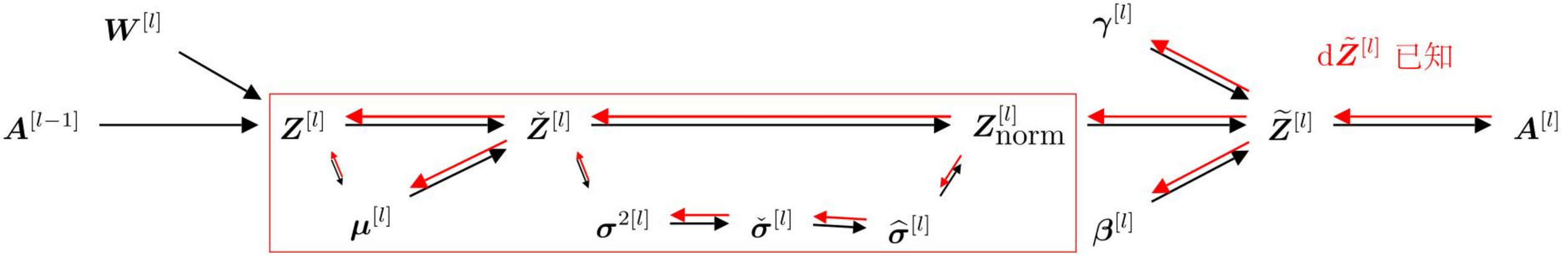
$$\begin{aligned}
\boldsymbol{\mu}^{[l]} &= m^{-1}(\boldsymbol{Z}^{[l]})^{\mathrm{T}}\mathbf{1} \in \mathbb{R}^{d^{[l]}\times 1}, \\
\boldsymbol{\sigma}^{2[l]} &= m^{-1}\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right)\circ\left(\boldsymbol{z}_i^{[l]} - \boldsymbol{\mu}^{[l]}\right)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

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

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

2．注意：对于批量归一化而言，偏置项 $\boldsymbol{b}^{[l]}$ 被均值中心化抵消



红色计算部分的后向传播过程为：

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

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

$$\mathrm{d}\boldsymbol{Z}_1^{[l]} = \mathrm{d}\check{\boldsymbol{Z}}^{[l]} \qquad \mathrm{d}\boldsymbol{\sigma}^{2[l]} = \mathrm{d}\check{\boldsymbol{\sigma}}^{[l]} \circ \hat{\boldsymbol{\sigma}}^{[l]}/2 \qquad \mathrm{d}\check{\boldsymbol{Z}}_1^{[l]} = \mathrm{d}\boldsymbol{Z}^{[l]}_{\text{norm}} \circ \left\{\mathbf{1}\left(\hat{\boldsymbol{\sigma}}^{[l]}\right)^{\mathrm{T}}\right\}$$

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

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

批量归一化的备注

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
2. 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变得更平滑）
   - 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能力

批量归一化的备注

1. 在验证和测试阶段，我们往往利用训练集上累计的均值和方差运行估计，即其对应的 runtime estimators，进行批量归一化处理
   - 在这两个阶段，样本点的数量可能较少，甚至只有一个
2. 更多的讨论，请参考 Santurkar（2018）等
3. 我们将在后续课程中，介绍其他的归一化方法
4. 为什么不在激活运算之后，进行批量归一化?

Dropout 与批量归一化的备注

1. 批量归一化通过更改每个特征对应的均值和方差，加快算法收敛，提高模型的泛化能力
2. Dropout 通过随机删除隐藏层的神经元，来提升模型的泛化能力
3. 当我们在 dropout 后实施批量归一化时，可能会遇到麻烦（陷阱一）
   - 在训练阶段，由于 dropout 随机丢弃神经元，其可能会改变未丢弃神经元的方差
   - 然而，批量归一化则在努力稳定神经元的方差
   - 此外，dropout 在模型的验证和测试阶段不再实施
   - 这可能会造成统计量偏移，导致训练阶段与推理阶段的统计量不匹配，进而降低推理性能（Li 等
4. 建议的实施顺序：

$$\text{线性变换} \rightarrow \text{批量归一化} \rightarrow \text{激活} \rightarrow \text{Dropout}$$

Dropout 与批量归一化的备注

1. 当我们同时实施批量归一化和 dropout 时，可能会造成过度正则化的风险（陷阱二）
   - 批量归一化具有一些正则化效果
   - 同时实施批量归一化和一个具有较高移除概率的 dropout，可能会导致欠拟合
2. 建议
   - 批量归一化 + 一个带有较小移除概率的 dropout
   - 只实施批量归一化

课程总结

1．深层网络训练困难的一大原因，是内部层输入分布，在训练中不断变化

2．基于白化操作的思想，批量归一化对每一个特征单独操作

3．批量归一化能稳定训练、支持更大学习率，并带来轻度正则化