

动手学深度学习 v2

李沐 · AWS



批量归一化

- 损失出现在最后，后面的层训练较快
- 数据在最底部
 - 底部的层训练较慢
 - 底部层一变化，所有都得跟着变
 - 最后的那些层需要重新学习多次
 - 导致收敛变慢
- 我们可以在学习底部层的时候避免变化顶部层吗？

损失



数据



批量归一化



- 固定小批量里面的均值和方差

$$\mu_B = \frac{1}{|B|} \sum_{i \in B} x_i \text{ and } \sigma_B^2 = \frac{1}{|B|} \sum_{i \in B} (x_i - \mu_B)^2 + \epsilon$$

然后再做额外的调整（可学习的参数）

$$x_{i+1} = \gamma \frac{x_i - \mu_B}{\sigma_B} + \beta$$

均值

方差



批量归一化层

- 可学习的参数为 γ 和 β
- 作用在
 - 全连接层和卷积层输出上，激活函数前
 - 全连接层和卷积层输入上
- 对全连接层，作用在特征维
- 对于卷积层，作用在通道维



批量归一化在做什么？

- 最初论文是想用它来减少内部协变量转移
- 后续有论文指出它可能就是通过每个小批量里加入噪音来控制模型复杂度

$$x_{i+1} = \gamma \frac{x_i - \hat{\mu}_B}{\hat{\sigma}_B} + \beta$$

随机偏移

随机缩放

- 因此没必要跟丢弃法混合使用



总结

- 批量归一化固定小批量中的均值和方差，然后学习出适合的偏移和缩放
- 可以加速收敛速度，但一般不改变模型精度