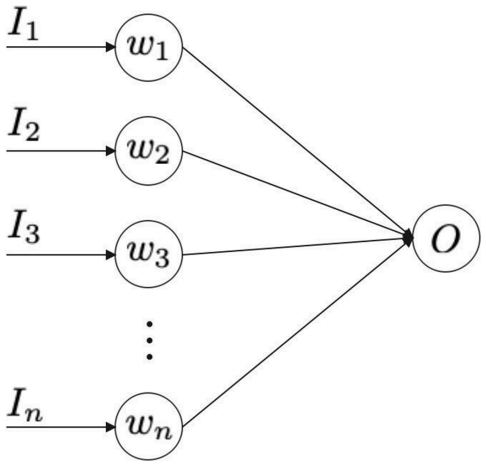
补充证明

更多相关证明和思路和讨论参考 **《深度学习：基础与概念》09 章 正则化** 笔记

1. dropout 等价于正则化



上图是一个线性的神经网络，它的输出是输入的加权和，表示为式 (1)。这里我们只考虑最简单的线性激活函数，这个原理也适用于非线性的激活函数，只是推导起来更加复杂。

$$O = \sum_{i}^{n} w_i I_i \tag{1}$$

对于无 dropout 的网络，它的误差可以表示为式 (2)，其中 t 是目标值。

$$E_N = \frac{1}{2}\left(t - \sum_{i=1}^{n} w_i' I_i\right)^2 \tag{2}$$

(2)式之所以使用 w' 是为了找到之后要介绍的加入Dropout的网络的关系，其中 $w' = pw$ 。那么(2)可以表示为式(3)。

$$EN = \frac{1}{2}\left(t - \sum_{i=1}^{n} p_i w_i I_i\right)^2 \tag{3}$$

它关于 w_i 的导数为式(4)。

$$\frac{\partial E_N}{\partial w_i} = -t p_i I_i + w_i p_i^2 I_i^2 + \sum_{j=1, j\neq i}^{n} w_j p_i p_j I_i I_j \tag{4}$$

当我们添加 dropout 之后，它的误差表示为式(5)。$\delta \sim \text{Bernoulli}(p)$ 是丢失率，它服从伯努利分布，即它有 p 的概率值为 1 ， $1-p$ 的概率值为 0。

$$E_D = \frac{1}{2}\left(t - \sum_{i=1}^{n} \delta_i w_i I_i\right)^2 \tag{5}$$

它关于 w_i 的导数表示为式(6)。

$$\partial E_D \partial w_i = -t \delta_i I_i + w_i \delta_i^2 I_i^2 + \sum_{j=1, j\neq i}^{n} w_j \delta_i \delta_j I_i I_j \tag{6}$$

因为 δ_i 是一个伯努利分布，我们对其求期望：

$$\begin{aligned} E\left[\frac{\partial E_D}{\partial w_i}\right] &= -tp_iI_i + w_ip_i^2I_i^2 + w_i\,\mathrm{Var}\,(\delta_i)I_i^2 + \sum_{j=1,j\neq i}^{n} w_jp_ip_jI_iI_j \\ &= \frac{\partial E_N}{\partial w_i} + w_i\,\mathrm{Var}\,(\delta_i)I_i^2 \\ &= \frac{\partial E_N}{\partial w_i} + w_ip_i\,(1-p_i)I_i^2 \end{aligned} \tag{7}$$

对比式(6)和式(7)我们可以看出，在 $w' = pw$ 的前提下，**带有 dropout 的网络的梯度的期望等价于带有正则的普通网络**。换句话说，dropout起到了正则的作用，正则项为 $w_ip_i\,(1-p_i)I_i^2$ 。

2. Batch normalization 是一种隐式正则化

① BN 的标准步骤：

1. 按 batch 求均值、方差

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} a_i, \quad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(a_i-\mu_B)^2$$

2. 标准化

$$\hat{a}_i = \frac{a_i - \mu_B}{\sigma_B}$$

3. 再做一层可学习的仿射变换

$$y_i = \gamma\hat{a}_i + \beta$$

> **Question**
>
> 之前一直以为 BN 只是做了归一化，这种仿射变换的必要性需要详细讨论

② 把 batch 统计量看作随机估计量

若假设 a_i 是从某个分布独立同分布采样得到的，则：

$$\mathbb{E}[\mu_B] = \mu, \qquad \mathbb{E}[\sigma_B^2] = \sigma^2$$

且有（中心极限定理型的近似）：

$$\mu_B = \mu + \delta_\mu, \quad \sigma_B = \sigma + \delta_\sigma$$

其中：

- $\delta_\mu, \delta_\sigma$ 是均值为 0 的随机扰动，方差约为 $O(\frac{1}{m})$
- batch 越小，δ 方差越大 → BN 噪声越强

把这些代回标准化公式：

$$\hat{a}_i = \frac{a_i - (\mu + \delta_\mu)}{\sigma + \delta_\sigma}$$

把分母做一次泰勒展开（假设 $|\delta_\sigma| \ll \sigma$）：

$$\frac{1}{\sigma+\delta_\sigma} \approx \frac{1}{\sigma} - \frac{\delta_\sigma}{\sigma^2}$$

于是：

$$\hat{a}_i \approx (a_i - \mu - \delta_\mu)\left(\frac{1}{\sigma} - \frac{\delta_\sigma}{\sigma^2}\right)$$

展开并按“确定部分 + 噪声部分”分组：

$$\hat{a}_i \approx \underbrace{\frac{a_i-\mu}{\sigma}}_{\text{理想标准化}} + \underbrace{\left(-\frac{\delta_\mu}{\sigma} - \frac{(a_i-\mu),\delta_\sigma}{\sigma^2}\right)}_{\text{BN 引入的噪声 } \xi_i}$$

记

$$\tilde{a}_i := \frac{a_i-\mu}{\sigma}, \quad \xi_i := -\frac{\delta_\mu}{\sigma} - \frac{(a_i-\mu),\delta_\sigma}{\sigma^2},$$

则可以写成简洁的形式：

$$\hat{a}_i = \tilde{a}_i + \xi_i$$

最后的输出

$$y_i = \gamma\hat{a}_i + \beta = \gamma\tilde{a}_i + \beta + \gamma\xi_i$$

定义：$\eta_i := \gamma\xi_i$，则

$$y_i = \underbrace{\gamma\tilde{a}_i + \beta}_{\text{“理想 BN 网络”的输出}} + \underbrace{\eta_i}_{\text{由有限 batch 统计量引入的噪声}}$$

总结： 真正训练时带 BN 的网络，可以看成是“一个确定网络 + 在每层输出加了一些特定结构的噪声”。

③ 反向传播对应隐式正则

由于前向有

$$y_i = y_i^{(0)} + \eta_i, \qquad y_i^{(0)} = \gamma\tilde{a}_i + \beta,$$

上游梯度变为

$$g_i = g_i^{(0)} + h_i\eta_i \quad (\text{一阶展开}).$$

把

$$g_i = g_i^{(0)} + h_i\eta_i, \qquad \hat{a}_i = \tilde{a}_i + \xi_i$$

代入 BN 的反向传播公式

$$\frac{\partial L}{\partial a_i} = \frac{1}{\sigma_B}\left(g_i - \frac{1}{m}\sum_j g_j\right) - \frac{\hat{a}_i}{m\sigma_B}\sum_j g_j\hat{a}_j,$$

并只保留关于 ξ_i, η_i 的一次小量，可得

$$\frac{\partial L}{\partial a_i} = G_i^{(0)} + \Delta_i,$$

其中

$$G_i^{(0)} = \frac{1}{\sigma}\left(g_i^{(0)} - \frac{1}{m}\sum_j g_j^{(0)}\right) - \frac{\tilde{a}_i}{m\sigma}\sum_j g_j^{(0)}\tilde{a}_j$$

是**没有 BN 噪声时的理想梯度。**

噪声诱导的额外梯度为

$$\begin{aligned}\Delta_i = \frac{1}{\sigma}\left(h_i\eta_i - \frac{1}{m}\sum_j h_j\eta_j\right) - \frac{\tilde{a}_i}{m\sigma}\sum_j g_j^{(0)}\xi_j \\ - \frac{\tilde{a}_i}{m\sigma}\sum_j h_j\eta_j + \frac{1}{m\sigma}\sum_j g_j^{(0)}\tilde{a}_j\,\xi_i + O(\xi\eta).\end{aligned}$$

再代入

$$\xi_i = -\frac{\delta_\mu}{\sigma} - \frac{(a_i-\mu)\delta_\sigma}{\sigma^2}, \qquad \eta_i = \gamma\xi_i,$$

即可得到

$$\boxed{\Delta_i = (\text{线性组合})\cdot\delta_\mu + (\text{线性组合})\cdot\delta_\sigma + O(\delta_\mu\delta_\sigma, \delta_\sigma^2)}$$

并且满足

$$\boxed{\mathbb{E}[\Delta_i] = 0, \qquad \mathrm{Var}(\Delta_i) = O\left(\frac{1}{m}\right)}$$

这意味着：

$$\boxed{\frac{\partial L}{\partial a_i} = \underbrace{G_i^{(0)}}_{\text{确定的 BN 梯度}} + \underbrace{\zeta_i}_{\text{由有限 batch 噪声产生的随机梯度}}}$$

其中 $\zeta_i = \Delta_i$。

这就是 BN 的隐式正则化来源：前向加噪声 + 反向也注入结构化梯度噪声，使优化动态更平滑、更具收缩性。

> **Thoughts**
>
> BN 带来的正则式源自统计学本身，是半结构化的，我们可以证明噪声大概的值，但是真正的噪声还是有随机性
>
> 但是在形式上，BN 和 dropout 的形式和带来的效果是基本一致的
>
> 所以 batch size 确实是一个需要调整的超参数，随机抽取的批次确实给模型训练带来了很多的可能性

Question

感觉和“SGD的隐式先验”很相似，都是来自于抽取的样本的随机性