

16. 假设用  $x \in B$  表示一个来自小批量  $B$  的输入, 批量归一化 BN 层的参数为: 拉伸参数 (scale)  $\gamma$  和偏移参数 (shift)  $\beta$ , 它们的形状与  $x$  相同。请写出经过批量归一化转换后的  $x$  表达式  $\text{BN}(x)$ 。

$$\text{BN}(x) = \frac{x - \mathbb{E}(x)}{\sqrt{\text{var}^2(x) + \epsilon}} \odot \gamma + \beta$$

17. 批归一化和 Dropout 一般不会同时使用, 请简述原因。

目标冲突: 批归一化的目的是调整激活层的输出, 使其保持一定的均值和方差, 从而加速训练过程并提高模型稳定性。Dropout 则是通过随机地关闭神经网络中的一部分神经元, 来防止模型对特定的训练数据过拟合。这两种技术的目标不同。批归一化希望保持激活值的分布稳定, 而 Dropout 则通过引入噪声来增加网络的泛化能力, 这在一定程度上是相互矛盾的。

内部协变量偏移: 批归一化通过规范化激活层的输出来减少内部协变量偏移。但是, 当使用 Dropout 时, 由于随机关闭神经元, 激活层的输出分布在每次迭代中都会发生变化, 这可能会破坏批归一化试图解决的问题。

训练动态: 当同时使用这两种技术时, 网络的训练动态可能会受到干扰。Dropout 会改变网络中信息的流动方式, 而批归一化依赖于这些信息的稳定分布来进行有效的规范化。这种干扰可能会导致训练过程不稳定, 甚至降低模型性能。

18. 简述 Dropout 能够防止过拟合的原因。

模拟集成学习: Dropout 通过在训练过程中随机关闭 (即设为零) 网络中的一部分神经元, 可以被看作是在训练每个小批量数据时都使用了一个略有不同的网络结构。这类似于训练多个不同的模型并进行集成学习。在测试时, 使用所有神经元 (通常会调整其输出以反映训练时的 Dropout 比例) 相当于对这些“模型”进行平均, 这有助于减少模型对特定训练样本的过度依赖, 提高其泛化能力。

减少复杂共适应性: 在没有 Dropout 的情况下, 网络的神经元可能会学会共同适应并对特定的训练数据过度拟合, 即它们共同调整它们的行为来记住训练数据的特定特征。使用 Dropout 意味着网络的神经元不能依赖于特定其他神经元的存在, 因为这些神经元可能在任何给定的训练步骤中被“丢弃”掉。因此, 每个神经元必须学习更加鲁棒的特征, 这些特征在各种不同的网络子集中都有用。

减少过度依赖: 在传统的神经网络中, 某些神经元可能对特定的特征或模式过度敏感, 导致过拟合。Dropout 强迫网络在缺少部分信息的情况下进行学习, 这减少了模型对任何单个神经元的依赖, 从而减少了过拟合的风险。

正则化效果: 从正则化的角度来看, Dropout 可以被视为一种增加模型的随机性、降低模型复杂度的方式。通过在训练期间引入噪声, Dropout 防止了网络对训练数据的“记忆效应”, 从而起到了类似于传统正则化技术 (如 L1 或 L2 正则化) 的作用。

19. 假设 Dropout 概率为  $p$ , 为了保证期望值不变, 则在测试时, 该层模型权重  $w$  应该变为?  $1 - p$

20. 设模型初始参数  $\theta_0$  为  $[0.1, 0.2]$ , 学习率为 0.1, 动量更新权重为 1, 若模型在两次训练过程中的梯度依次为  $g_0 = [2, -1]$ ,  $g_1 = [1, -2]$ , 求两次训练后模型的参数  $\theta_2$ .  $[-0.4, 0.6]$