

15.1 随机失活 (dropout) 是一个训练神经网络时很有用的技术, 它由 Srivastava 等人在 JMLR 发表的论文“Dropout: A Simple Way to Prevent Neural Networks from Overfitting”中提出 [222]^⑤. 仔细阅读该论文并回答如下问题 (请将每问的答案组织为一个简洁的语句).

- (a) 随机失活在训练时如何操作?
- (b) 随机失活在测试时如何操作?
- (c) 随机失活有什么好处?
- (d) 为什么随机失活可以得到这个好处?

(a) 在训练过程中, Dropout 的实现是让神经元以超参数 p 的概率停止工作或者激活被置为 0

(b) 在测试过程中, 不进行随机失活, 而是将 Dropout 的参数 p 乘以输出。

(c) 随机失活通常降低了神经网络的计算开销。

(d) 因为归零操作可以得到稀疏矩阵。

15.2 VGG16 CNN 模型 (也称为 VGG-Verydeep-16) 由 Karen Simonyan 和 Andrew Zisserman 在 arXiv 预印服务器上公布的论文“Very Deep Convolutional Networks for Large-Scale

Image Recognition”中提出 [219]^⑥. 此外, GoogLeNet 模型由 Szegedy 等人在 arXiv 预印服务器上公布的论文“Going Deeper with Convolutions”中提出 [226]^⑦. 这两篇论文几乎同时公开, 并共享一些相似的思路. 仔细阅读两篇论文并回答如下问题 (请将每问的答案组织为一个简洁的语句).

- (a) 为什么它们使用小的卷积核 (主要是 3×3) 而不是大的卷积核?
- (b) 为什么这两个网络都相当深 (即有许多层, 大概 20 左右)?
- (c) 什么困难是由大的深度造成的? 在这两个网络中是如何解决的?

(a) 因为这样可以减少参数个数。

(b) 因为增加深度可以提高神经网络的性能。

(c) (1) 容易过拟合; (2) 计算需要的资源增加, 消耗内存, 算力等资源。

VGG16 的网络的解决办法为加深结构全部使用 ReLU 激活函数, 以此来提升非线性变化的能力。

GoogLeNet 团队提出了 Inception 的新型网络结构，来搭建一个稀疏性、高计算性能的网络结构。

15.3 批量规范化 (Batch Normalization, BN) 是另一个在训练深度神经网络时很有用的技术，它由 Sergey Ioffe 和 Christian Szegedy 在 ICML 2015 发表的论文“Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”中提出 [121]。仔细阅读该论文并回答如下问题 (请将每问的答案组织为一个简洁的语句)。

- (a) 什么是内部协变量转移 (internal covariate shift)?
- (b) BN 是如何处理这个问题的?
- (c) BN 在卷积层内如何操作?
- (d) BN 有什么好处?

(a) 每层输入的分布在训练过程中会发生变化，因为前面的层的参数会发生变化。通过要求较低的学习率和仔细的参数初始化减慢了训练，并且使具有饱和和非线性的模型训练起来非常困难。我们将这种现象称为内部协变量转移。

(b) 它通过对神经网络的中间层进行逐层归一化，让每一个中间层输入的分布保持稳定，即保持同一分布。

(c) (1) 求数据均值

(2) 求数据方差

(3) 数据进行标准化

(4) 训练参数 γ , β

(5) 输出 y 通过 γ 与 β 的线性变换得到新的值

(d) (1) 可以使用更高的学习率。

(2) 移除或使用较低的 Dropout。

(3) 降低 L2 权重衰减系数。

(4) 取消 Local Response Normalization 层。

(5) 减少图像扭曲的使用。

15.4 ResNet 是一个非常深神经网络的学习技术, 由 He 等人在 CVPR 2016 发表的论文“Deep Residual Learning for Image Recognition”中提出 [109][®]. 仔细阅读该论文并回答如下问题 (请将每问的答案组织为一个简洁的语句).

- (a) 尽管 VGG16 和 GoogLeNet 在训练大约 20~30 层深的网络时遇到了困难, 是什么使得 ResNet 可以训练深如 1000 层的网络?
- (b) VGG16 是一个前馈网络, 其中每一层只有一个输入和一个输出. 而 GoogLeNet 和 ResNet 是有向无环图结构 (Directed Acyclic Graph, DAG), 只要网络结构中的数据流没有构成一个环, 一层能够有多个输入和多个输出. DAG 和前馈结构相比有什么好处?
- (c) VGG16 有两个全连接层 (fc6 和 fc7), 而 ResNet 和 GoogLeNet 没有全连接层 (除了用于分类的最后一层). 什么被用来取代 FC? 这有什么好处?

(a) 让神经网络某些层跳过下一层神经元的连接, 弱化每层之间的强联系。

(b) 能更快的向前传播数据或向后传播梯度。

(c) 使用了全局平均池化层, 可以减少大量参数, 提高效率。

15.5 AlexNet 指在 ILSVRC 竞赛数据上训练的深度卷积神经网络, 这是在计算机视觉任务中深度 CNN 的一个开创性工作. AlexNet 的技术细节在论文“ImageNet Classification with Deep Convolutional Neural Networks”中给出, 这是由 Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E. Hinton 在 NIPS 25 发表的论文 [136][®]. 它提出 ReLU 激活函数, 并创造性地使用 GPU 来加速计算. 仔细阅读该论文并回答如下问题 (请将每问的答案组织为一个简洁的语句).

- (a) 表述你怎么理解 ReLU 帮助 AlexNet 取得成功? GPU 又是如何助力 AlexNet 的?
- (b) 使用多个网络的预测的平均会降低错误率. 为什么?
- (c) 随机失活技术是在哪里被应用的? 它是如何提供帮助的? 使用随机失活有什么代价?
- (d) AlexNet 中有多少参数? 为什么数据集大小 (120 万) 对 AlexNet 的成功很重要?

(a) 成功使用 ReLU 作为 CNN 的激活函数, 并验证其效果在较深的网络超过了 Sigmoid, 成功解决了 Sigmoid。

阵运算

(b) 避免模型发生过拟合。

(c) 在训练时使用。能够防止过拟合。

(d) 约 6000 万，因为数据量越多，越能避免过拟合的发生。