理解深度学习

许志钦

2023 年 9 月 13 日

目录

Chapter 1

泛化问题

1.1 引言

在前一章中，我们将神经网络的误差分成三类：逼近误差、泛化误差、优化误差。逼近误差反映的是模型的表达能力，一般当我们使用的模型比较大时，逼近误差可以认为非常小。优化误差是通过寻找合适的优化方法将训练集上的误差降到最低，也就是常说的寻找全局最小点。泛化误差是带有一定的运气成分的，因为我们希望从已经的数据中学习到的模型能够应用到未知的数据。原则上，只有对未知数据有一定假设，我们才能谈泛化。实际问题之所有问题是因为实际数据通常非常高维和复杂，我们很难找到合适的方法描述实际数据的特征。在这节中，我们简要的回顾传统的学习理论怎么研究泛化，深度学习的泛化有什么重要问题，以及哪些因素可能会影响深度学习的泛化。

1.2 传统学习理论对于泛化误差的理解

在继续推进关于神经网络泛化性理解之前，让我们先一起来看看传统机器学习理论是怎么理解泛化误差的。这能为我们理解神经网络的泛化误差提供一些基础性的观点。

在传统机器学习理论中，科学家们建立了一套成熟的理论体系与研究方式来估算一个算法的泛化误差，我们将通过下面的例子来理解传统的学习理论。

1) 龙格现象 (Runge's Phenomenon)：在回归问题中，使用复杂度更高的模型，往往会带来剧烈的震荡，（正如下图所示。）

2) 欠拟合、过拟合现象：（加过拟合和欠拟合的图）

传统机器学习理论用复杂度 (比如 VC-dimension，Rademacher Complexity) 来衡量一个

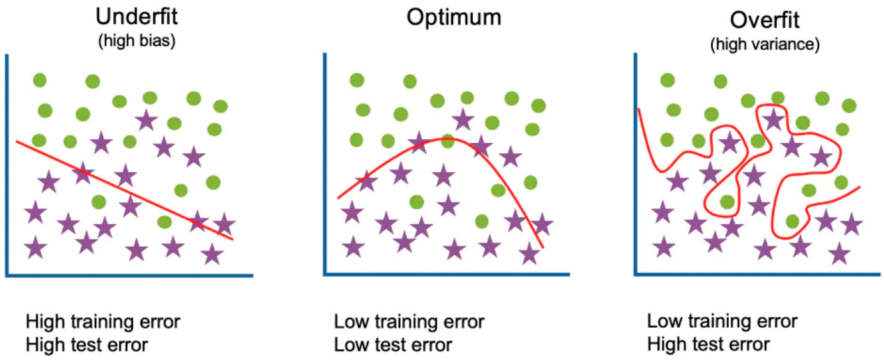


图 1.1: 常见的拟合问题。

算法所使用模型的函数空间的大小。复杂度可以简单地理解为模型参数的多少，模型参数越多，则模型复杂度越高，模型所处的函数空间也就越大。显然，如果复杂度太小，该算法可能连逼近训练集也做不到。而当复杂度过高的时候，模型又会带来很多不必要的成分。这些成分的来源是多方面的，比如在初始化的时候，由于模型可的高度复杂性，初始的参数使得模型有很多不同的成分比如很多高频振荡的信息，但由于训练数据并没有这么高频的信息去更正初始的这些成分，使得这些成分可能会保留到最后，或者由于训练样本在采样时有一些噪音，但模型却又足够复杂，这导致其可以很容易地也将这些噪音拟合好，进而过拟合。

定量上，理论在考虑最差情况下 Shalev-Shwartz & Ben-David (2014)，通常可以得到如下估计

$$\text{测试误差} \leq \text{训练误差} + \text{模型复杂度} + \text{小量} \tag{1.1}$$

所以传统的学习理论给了一个很重要的启发，即，选择模型的时候不要选择过于复杂的模型。这个启发在科学研究中也是很常见的，也深刻地影响了很多人做研究的方式，比如冯 • 诺依曼说过：“给我四个参数我可以拟合一头大象，再给一个，我可以让大象动起来。”(Dyson 2004)。

科学领域中还有一条准则为奥卡姆剃刀（“Ocam's Razor”），讲的是“如无必要，勿增实体”，即，如果两个理论都能解释同一个现象，那我们会倾向于选择假设少且更为简单的理论。

经常有人在使用神经网络后，得出效果很差的结论，我发现他们用的神经网络仅有一两层，且神经元的数目也就两三个，一问原因，他们解释说参数过多容易导致过拟合。这正是受到传统学习理论的影响。这里涉及到神经网络与传统学习理论之间一个表面的矛盾，我们将在下一节详细讲。

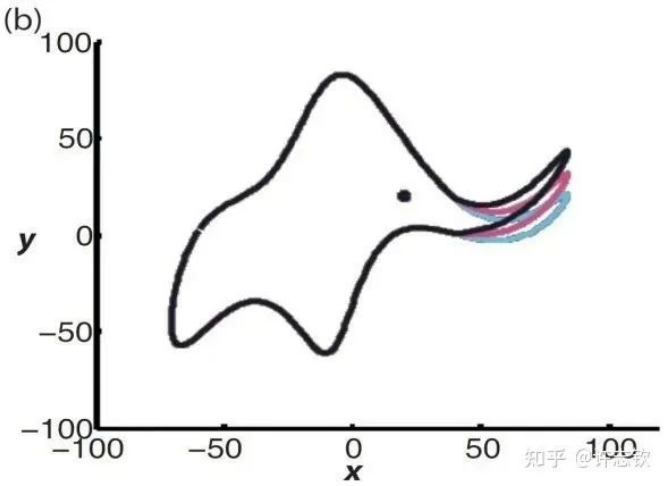


图 1.2: 冯 • 诺依曼说过：“给我四个参数我可以拟合一头大象，再给一个，我可以让大象动起来。”(Dyson 2004)

1.3 神经网络中泛化误差迷因

1995 年 Leo Breiman 在总结 NIPS 文章的时候Breiman (1995)，对深度学习提出了一些至今仍然很重要的问题

- 为什么过参数化 (Over-parameterized) 的神经网络对于给定数据不会过拟合？
- 神经网络的有效参数数目是多少？
- 为什么反向传播 (Backpropergation) 算法不会让模型进入一个误差大的局部最小点？
- 应该在什么时候停止训练？

第一个问题谈的是为什么过参数化的神经网络在很多实际问题中不过拟合，也就是参数数目大于训练样本的数目的情况。这个问题有趣的点在于它看起来与传统的学习理论有显著的矛盾，传统的学习理论强调参数少或者说模型的低复杂度是保证算法泛化性的重要因素，但神经网络却 “鼓励” 用大量参数的模型，并且不过拟合。2017 年 ICLR 的论文，zhang et al 用一些列的实验再次将这个问题带进了大家的视线Zhang et al. (2016)。以下面这个实验为例我们来看一下实际训练的情况。

图表1.3中的模型有 160 万个参数，数量上远大于 CIFAR10 中 6 万个数据。正如最后一行所展示的，即使打乱所有图片的标签，神经网络仍可以完全拟合这些训练数据点，此时测试集的准确率为 9.78%，接近 10% 的原因是数据集是由 10 类样本构成的，这个例子说明神经网

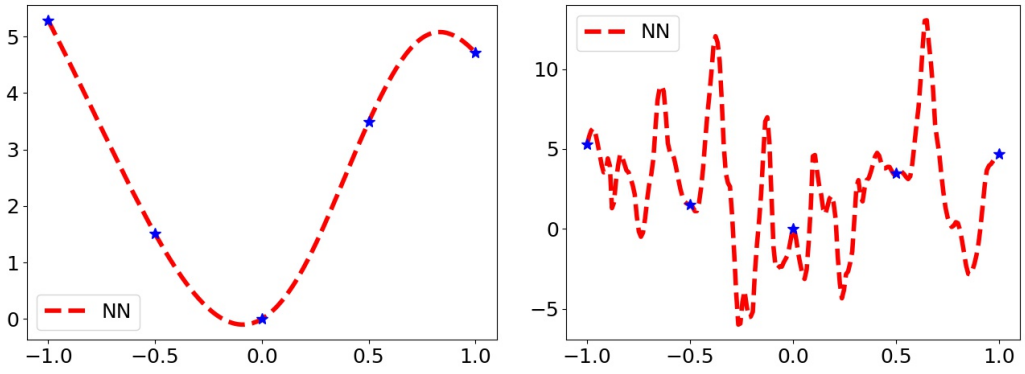
Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset. Performance with and without data augmentation and weight decay are compared. The results of fitting random labels are also included.

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78

图 1.3: 大模型不过拟合的例子。图表来源于Zhang et al. (2016)

络的表达能力非常强。而第一到第四行的结果表明，无论是否使用训练技巧，神经网络都能准确分类训练集且并没有明显的过拟合。

这个高维的例子看起来非常抽象，让我们用一个一维的插值实验来进一步理解这个现象。



(a) 小初始化。 (b) 大初始化。

图 1.4: 一维拟合例子的结果。用四个隐藏层的神经网络拟合蓝色星点数据，每层 500 个神经元，激活函数为 Tanh。两张图的差别在于参数初始化的高斯分布的方差不一致。

图1.4的例子中，不同宽度及深度的神经网络被用来拟合五个蓝色数据点。对于大部分正常 (Gaussian、Lecun、Xavier 等) 的初始化，如左图所示，神经网络均能用相对平坦的解来拟合训练点 (虚线)。显然在这些例子中，神经网络并没有表现出龙格现象中的振荡。我们会倾向于说没有剧烈振荡的解是“好”的解。

当然，这个一维的实验里还存在几个问题。首先，神经网络是不是没有办法用复杂振荡的曲线来拟合训练集？于是我们改变神经网络的初始化，将神经网络的参数初始化改为较大的值，如图1.4b所示，可以看到此时神经网络用非常振荡的曲线拟合给定的点，所以神经网络的

结构是有能力用复杂的曲线拟合训练点的。其次，有人会争论这里没有泛化性能大的概念，要是目标函数刚好是非常振荡的曲线，那么前面用正常初始化得到的解的泛化性能就不好了，反而是那个振荡的解可能有更好的泛化性。这里争议的核心是 “泛化” 的概念。若我们仅有训练数据点，当拟合好的时候，没有争议的是目标函数和神经网络均会通过这几个训练点，但谁也无法说明在非给定的位置上，目标函数和神经网络各是什么值，或者说它们是什么值都是可以的。既然这样，那我们谈论泛化性的时候就必须指明是在什么数据上讨论的。尽管我们无法准确描述真实数据具有什么特点，但我们会觉得振荡厉害的解泛化性好的概率很低。

无论是一维还是高维的例子，都可以看出常用的神经网络似乎有大量参数，但却不像传统学习理论预测的那样，呈现出很差的泛化或有剧烈的振荡。2017 年以后，这个问题成为理论研究的热点之一，这也是本书将着重讨论的一个问题。

这个关于泛化性的谜团很自然的引出 Breiman 的第二个问题，神经网络的有效参数数目是多少？打个极端的比方，如果两个神经元的权重一样，那它们实际上用一个神经元就够了，或者说有一些神经元的参数为 0，那么这些神经元可以被直接拿掉。因此研究神经网络的有效参数比研究它表面的参数数目更有意义。而有效参数的研究涉及到神经网络的参数演化，损失函数的极值点分布、宽网络与窄网络、深网络与浅网络的联系，这部分我们将在本书后面的部分深入研究。

Leo Breimen 的第三个问题是非凸优化问题中常见的问题。凸函数的特点是所有的极值点均为最小点，但显然神经网络的优化问题是高度非凸问题，它的损失景观可能存在大量的鞍点、局部最小点等。为什么简单的梯度下降的变种方法就可以找到训练误差和泛化误差都很小的点？

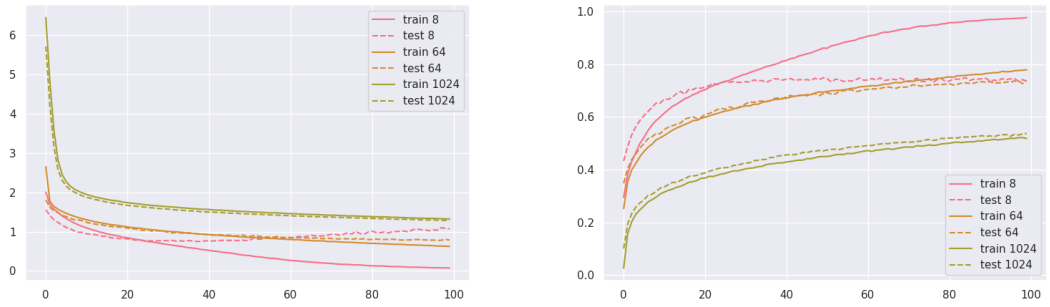
Leo Breimen 的第四个问题指出一个重要的事实，在实际训练中，我们并不寻求优化误差或者说训练误差达到最小，而是追求泛化误差最小，而泛化误差最小的情况往往不是优化误差最小的情况。如果解释这个现象，以及指导实际训练在何时停止训练，我们将在频率原则部分作探讨。

1.4 影响神经网络泛化误差的因素

讲了这么多关于影响算法模型泛化性能的分析与例子，那对于神经网络而言，到底什么会影响它的泛化性能呢？我们接下来将从... 这个几个角度出发，逐一讲解其与神经网络泛化性之间的关系。

我们使用 CIFAR10 数据集进行实验，分别考虑了 batch size、dropout、网络结构、网络深度、优化器等对训练结果的影响。同时，我们还使用神经网络拟合不同频率的目标函数 $\sin(vx)$ 来说明目标函数的频率对泛化性的影响。

1.4.1 使用随机梯度下降算法进行训练



(a) loss (b) accuracy

图 1.5: 使用随机梯度下降算法进行训练，考虑不同 batch size 对泛化性的影响

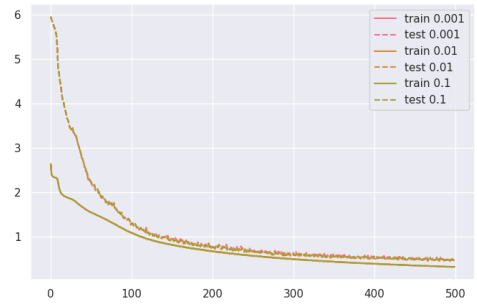
如图1.5所示,我们使用 resnet18 对 CIFAR10 数据进行分类预测,我们使用相同的 dropout、网络结构、网络深度、优化器、学习率，并使用不同的 batch size 进行训练。从图中，我们可以看出使用不同的 batch size，模型在训练集和测试集上的效果有着很大的差异，也即对模型的泛化性有一定的影响。

1.4.2 算法训练中使用 Dropout

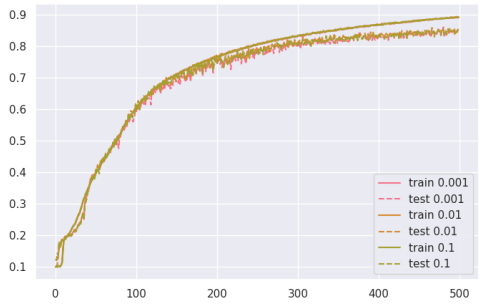
考虑到 Batch Normalization 对 dropout 的影响，因此，在说明 dropout 对模型泛化性的影响的实验中，我们选用 VGG16 进行实验，同样，我们保持其他影响训练结果的实验设置一致，仅改变 dropout 的比例（我们在 VGG16 网络中的激活函数处增加 Dropout 层，不改变其他网络结构）。如图1.6

1.4.3 神经网络的结构

为了说明网络结构对泛化性的影响，我们对比了三种常见的网络结构进行实验，简单的卷积神经网络，深度神经网络，在神经网络中引入残差结构，选取了常见的网络两层的卷积网络，VGG16，resnet18 进行实验。如图1.7，我们使用两层的卷积神经网络、VGG16，renet18 网络对 CIFAR10 数据进行分类预测。不同的网络在训练过程中，损失和准确率在训练集和测试集上有着不同的表现。

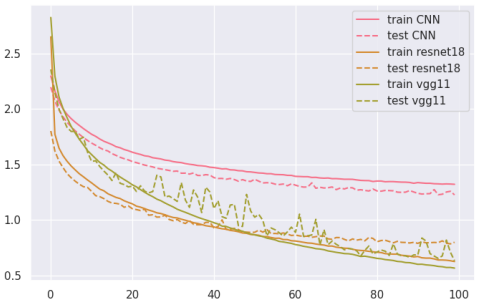


(a) loss

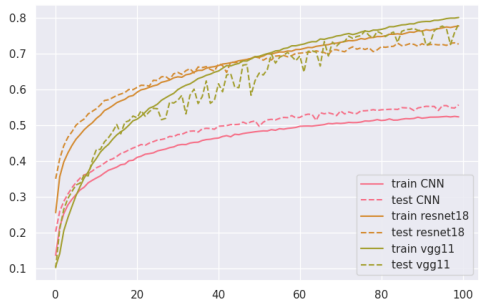


(b) accuracy

图 1.6: 使用随机梯度下降算法进行训练，考虑不同 dropout 对泛化性的影响



(a) loss

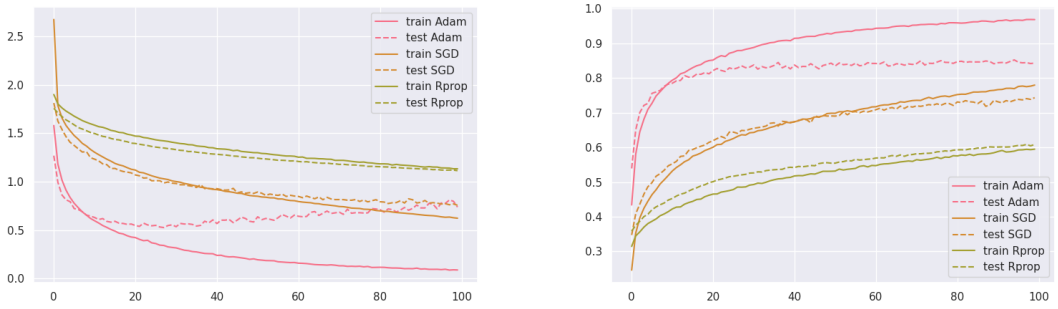


(b) accuracy

图 1.7: 使用随机梯度下降算法进行训练，考虑不同网络对泛化性的影响

1.4.4 优化器

如图1.8，我们使用 Adam、SGD、Rprop 优化器训练 resnet18。不同的优化器训练得到的模型在损失和准确率在训练集和测试集上有着不同的表现。



(a) loss (b) accuracy

图 1.8: 使用随机梯度下降算法进行训练，考虑不同优化器对泛化性的影响

1.4.5 神经网络的频率

我们还使用神经网络拟合不同频率的目标函数 sin(vx) 来说明目标函数的频率对泛化性的影响。

1.5 通过隐式偏好研究泛化问题

万有逼近定理指出两层的神经网络就可以表现出非常强大的逼近能力Cybenko (1989)，但令人感兴趣的常常不是在训练点上的逼近程度，而是神经网络对没见过的数据的预测能力。

任何一个学习方法在给定训练集、训练方法、初始参数等一系列条件后，均会得到一个确定的函数。一般来说，在测试点上，真实的标签可能是任意的。对两个不同的目标函数，若他们在采样点上的值是一样的，那么给定的训练方法若能精确预测好其中一个函数，则必然不能预测好另一个函数。更一般地，对于任意一种模型，我们总是可以找到一类数据使得它没办法基于给定的训练集预测好测试集。这些描述可以更准确地总结为没有免费的午餐定理 (No Free Lunch)。由于对所有可能函数的相互补偿，最优化算法的性能总是等价的。因此，简单地声称某个算法比另外一个算法好是一种不严谨的说法。一般情况下，我们说某个算法比较好是基于一种双方都默认的数据类型上的。

为了研究一个方法的泛化性，我们有必要把研究分成两个部分，一个是算法的特征，另一个是数据的特征。如果算法和数据的特征是一致的，那算法的泛化性能就会好，若不一致，则泛化性能就会较差。基于复杂度的分析研究泛化性，似乎只需要关心算法的模型的特征，而不用考虑数据的类型，但实际上这类分析是考虑了最差的情况，因此，基于复杂度的分析结果常常很难应用到实际的问题Jiang et al. (2019)。

算法特征和数据特征匹配的想法在传统的优化问题中常被使用。例如，图像恢复问题，真实图像是未知的，对于没有被污染的像素点，它们值是确定的，但对于被污染的数据点，如何从干净的数据点来推测它们是不确定的。过去的几十年，科学家们寻找各类正则化技术，比如限制导数的一范数或者二范数等。不同的正则化项对应的就是算法的特征，而算法能不能起到好的作用，需要看它是否反应了真实图像的特征。比如最小化一范数（total variation），得到的图像允许有大的突变（边界），但要求这些边界在整个图像中是比较稀疏的。这样的要求和自然图像的特征很靠近，因此也成为非常流行的方法。

对于过参数化的神经网络，同样是存在多解的问题。如何从多解中找到泛化好的解是重要的问题。实际的训练通过给调整各类超参、选训练方法、初始化等找到一个确定的解。和传统的优化问题比起来，神经网络的问题在没有显式的加正则项的情况下，能找到一类泛化好的解。一个自然的思路是能否把这些需要调整的超参、训练方法、初始化等转化成一个等价的正则项。这个正则项体现了神经网络的隐式偏好，也就容易理解算法的特征了。

本书的一大部分内容就是要探索神经网络的隐式偏好。简单的总结如下。在所有可行解中，神经网络会偏好寻找低频的解。在低频的解中，当初始化是在线性区域，神经网络会找距离初始点最近的解；当初始化在非线性区域时，神经网络会偏好于找参数凝聚的解。具体的解释将在后面的章节中展开。

参考文献

Breiman, L. (1995), 'Reflections after refereeing papers for nips', *The Mathematics of Generalization* **XX**, 11–15.

Cybenko, G. (1989), 'Approximation by superpositions of a sigmoidal function', *Mathematics of control, signals and systems* **2**(4), 303–314.

Dyson, F. (2004), 'A meeting with Enrico Fermi', *Nature* **427**(6972), 297–297.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D. & Bengio, S. (2019), 'Fantastic generalization measures and where to find them', *arXiv preprint arXiv:1912.02178* .

Shalev-Shwartz, S. & Ben-David, S. (2014), *Understanding machine learning: From theory to algorithms*, Cambridge university press.

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2016), 'Understanding deep learning requires rethinking generalization', *arXiv preprint arXiv:1611.03530* .