# 摘要

尽管体积庞大，但成功的深部人工神经网络可以显示出训练和测试之间的差异非常小。常规的智慧将小的泛化误差归因于模型族的特性，或者归因于训练过程中使用的正则化技术。通过广泛的系统实验，我们展示了这些传统的方法如何不能解释为什么大型神经网络在实践中推广得很好。

具体来说，我们的实验建立了最先进的卷积网络对于用随机梯度法训练的图像分类，很容易对训练数据进行随机标记。这种现象在质量上不受影响。

通过显式正则化，即使我们用完全非结构化随机噪声替换真实图像也会发生这种情况。我们证实了这些实验结果。

理论结构表明，两个神经网络在参数数目一到，就具有良好的有限样本表达能力。

超出了数据点的数量，这在实践中通常是如此。我们通过与传统模型的比较来解释我们的实验结果。

# 1介绍

深层人工神经网络通常比他们接受过培训的样品。尽管如此，其中一些模型显示出非常小的泛化误差，即“训练误差”和“测试误差”之间的差异。同时，它是当然，很容易提出概括性较差的自然模型体系结构。那么它是什么这将神经网络与非神经网络区别开来。令人满意的回答对于这个问题，不仅有助于使神经网络更易于理解，而且也可能使模型体系结构设计更具原则性和可靠性。

为了回答这个问题，统计学习理论提出了许多不同的复杂性能够控制泛化误差的度量。这些包括VC尺寸（Vapnik，1998年）、Rademacher复杂性（Bartlett和Mendelson，2003年）和均匀稳定性（Mukherjee等人，2002；Bousquet&Elisseeff，2002；Poggio等人，2004）。而且，当参数很大，理论表明需要某种形式的正则化来确保较小的泛化错误。正则化也可能是隐式的，就像早期停止的情况一样。

## 1.1我们的工作

在这项工作中，我们通过表明它无法区分具有根本不同的泛化性能的不同神经网络来对传统的泛化观点进行问题化。

**随机化测试**

我们的方法论的核心是来自非参数统计的着名随机化测试的变体（Edgington＆Onghena，2007）。 在第一组实验中，

我们在数据副本上训练几个标准体系结构，其中真正的标签被随机标签替换。 我们的核心发现可归纳为：

深度神经网络很容易适应随机标签。

更确切地说，当对真实数据的完全随机标记进行训练时，神经网络实现0训练误差。 当然，测试错误并不比随机机会好，因为训练标签和测试标签之间没有相关性。 换句话说，通过单独随机化标签，我们可以强制模型的泛化误差在不改变模型，大小，超参数或优化器的情况下大幅跳跃。 我们为在CIFAR10和ImageNet分类基准上训练的几种不同标准体系结构建立了这一事实。 虽然简单易懂，但从统计学习的角度来看，这一观察结果具有深远意义：

1.神经网络的有效容量足以记忆整个数据集。

2.即使是随机标签上的优化也很容易。 事实上，与真实标签上的培训相比，培训时间仅增加了很小的常数因素。

3.随机化标签只是一种数据转换，使学习问题的所有其他属性保持不变。

在第一组实验中，我们还用完全随机的像素（例如，高斯噪声）替换真实图像，并观察卷积神经网络继续适合数据

没有训练错误。这表明，尽管它们的结构，卷积神经网络可以适应随机噪声。我们进一步改变了随机化的数量，在两者之间平滑插值

没有噪音和完全噪音的情况。这导致一系列中间学习问题，其中标签中仍存在一定程度的信号。我们观察到了稳定的恶化

随着我们增加噪声水平的泛化误差。这表明神经网络能够捕获数据中的剩余信号，同时使用蛮力拟合噪声部分。

我们将在下面进一步详细讨论这些观察结果如何排除所有VC维度，Rademacher复杂性和统一稳定性，作为对泛化性能的可能解释。

最先进的神经网络。

**显性正规化的作用**

如果模型架构本身不是一个足够的正则化器，那么仍然需要了解明确的正则化有多大帮助。 我们表明，正则化的显式形式，如权重衰减，辍学和数据增加，不能充分解释神经网络的泛化误差。 换句话说：

显式正则化可以改善泛化性能，但既不必要也不足以控制泛化误差。

与经典的凸经验风险最小化相比，显性正则化是排除琐碎解的必要条件，我们发现正则化在深度学习中起着相当不同的作用。 它似乎更像是一个调整参数，通常有助于改善模型的最终测试误差，但缺少所有正则化并不一定意味着不良的泛化误差。 正如Krizhevsky等人报道的那样。 （2012），L2正则化（权重衰减）有时甚至有助于优化，说明其在深度学习中的理解不足。

**有限的样本表达性**

我们用理论结构补充我们的经验观察，表明一般大的神经网络可以表达训练数据的任何标记。更正式地说，我们展示了一个非常简单的双层ReLU网络，其中p = 2n + d个参数可以表示任何尺寸为n维的样本的任何标记。 Livni等人的先前建筑。 （2014）用更多的参数实现了类似的结果，即O（dn）。虽然我们的深度2网络不可避免地具有大的宽度，但我们也可以提出深度k网络，其中每个层仅具有O（n / k）参数。而先前的表达性结果集中于神经网络在整个域上可以表示的功能。 ，我们关注于有限样本的神经网络的表达性。与功能空间中现有的深度分离（Delalleau＆Bengio，2011; Eldan＆Shamir，2016; Telgarsky，2016; Cohen＆Shashua，2016）相比，我们的结果表明即使是线性尺寸的深度2网络也已经代表任何标签的训练数据。

**隐式正规化的作用**

虽然辍学和体重衰减这样的明确的正规化因素可能不是推广的必要条件，但肯定的是并非所有符合训练数据的模型都能很好地推广。 实际上，在神经网络中，我们几乎总是选择我们的模型作为运行随机梯度下降的输出。 对线性模型很有吸引力，我们分析了SGD如何充当隐式正则化器。 对于线性模型，SGD始终收敛于具有小范数的解决方案。 因此，算法本身隐含地使解决方案正规化。 实际上，我们在小数据集上显示，即使是高斯核方法也可以很好地推广而没有正则化。 虽然这并不能解释为什么某些体系结构比其他体系结构更好地概括，但它确实表明需要进行更多调查才能准确理解使用SGD训练的模型所继承的属性。

## 1.2 相关工作

哈特等人。 （2016）给出了随机梯度下降训练模型在梯度下降步数方面的泛化误差的上界。 他们的分析贯穿了统一稳定的概念（Bousquet＆Elisseeff，2002）。 正如我们在这项工作中指出的那样，学习算法的统一稳定性与训练数据的标记无关。 因此，该概念不足以区分在真实标签上训练的模型（小的泛化误差）和在随机标签上训练的模型（高泛化误差）。 这也突出了为何分析Hardt等人。 （2016）非凸优化是相当悲观的，只允许很少的数据传递。 我们的研究结果表明，即使是经验训练神经网络，对于数据的多次传递也不是一致的。 因此，需要较弱的稳定性概念才能沿着这个方向取得进一步进展。

关于神经网络的表征能力已有很多工作，从多层感知器的通用逼近定理开始（Cybenko，1989; Mhaskar，1993; Delalleau＆Bengio，2011; Mhaskar＆Poggio，2016; Eldan＆Shamir，2016; Telgarsky，2016; Cohen＆Shashua，2016）。所有这些结果都处于人口水平，表征某些神经网络家族在整个领域中可以表达哪些数学函数。相反，我们研究了大小为n的有限样本的神经网络的表示能力。这导致了一个非常简单的证据，即使O（n）大小的双层感知器具有通用的有限样本表达性。

巴特利特（1998）证明了在每个节点的权重的L1范数方面，多边形感知器的脂肪破碎维度与S形激活的界限。这个重要的结果给了一个

神经网络的泛化界限与网络大小无关。但是，对于RELU网络，“1-norm”不再提供信息。这导致了一个问题，即是否存在一种不同形式的容量控制，它限制了大型神经网络的泛化误差。 Neyshabur等人在一篇发人深省的着作中提到了这个问题。 （2014），通过实验证明网络规模不是神经网络容量控制的主要形式。类比矩阵分解说明了隐式正则化的重要性。

# 2神经网络的有效容量

我们的目标是了解前馈神经网络的有效模型容量。往为了实现这一目标，我们选择了一种受非参数随机化测试启发的方法。 特别，我们采用候选架构并在真实数据和数据副本上对其进行训练其中真正的标签被随机标签取代。 在第二种情况下，不再有任何

实例和类标签之间的关系。 结果，学习是不可能的。 直觉表明这种不可能性应该在训练期间清楚地表现出来，例如不通过训练大幅收敛或放缓。 令我们惊讶的是，培训过程的几个属性对于多个标准建筑，很大程度上不受标签转换的影响。 这个提出了一个概念上的挑 无论我们有什么理由期待一个小概括错误开始必须不再适用于随机标签的情况。

为了进一步了解这种现象，我们尝试了不同级别的随机化探索无标签噪音和完全损坏的标签之间的连续统一体。 我们也试试输入的不同随机化（而不是标签），得出相同的一般结论。实验在两个图像分类数据集CIFAR10数据集（Krizhevsky＆Hinton 2009）和ImageNet（Russakovsky等，2015）ILSVRC 2012数据集。 我们测试了ImageNet上的初始V3（Szegedy等，2016）架构和较小版本的Inception，lexnet（Krizhevsky等，2012），以及CIFAR10上的MLP。 请参阅附录中的A部分实验设置的更多细节。

## 2.1适合随机标签和像素

我们通过以下标签和输入图像的修改来运行我们的实验：

•真实标签：未经修改的原始数据集。

•部分损坏的标签：独立于概率p，每个图像的标签是作为一个统一的随机类被破坏。

•随机标签：所有标签都替换为随机标签。

•混洗像素：选择像素的随机排列，然后将相同的排列应用于训练和测试集中的所有图像。

•随机像素：独立地对每个图像应用不同的随机排列。

•高斯：高斯分布（具有与原始图像匹配的均值和方差）数据集）用于为每个图像生成随机像素。

令人惊讶的是，具有未改变的超参数设置的随机梯度下降可以优化即使随机标签完全破坏了，也可以完美地适应随机标签图像和标签之间的关系。我们通过改组进一步打破了图像的结构图像像素，甚至从高斯分布中完全重新采样随机像素。但我们测试的网络仍然能够适应。图1a显示了各种设置下CIFAR10数据集上的Inception模型的学习曲线。我们希望目标函数在随机标签上开始减少需要更长的时间因为最初每个训练样本的标签分配是不相关的。因此，大预测误差被反向传播以使参数更新产生大的梯度。然而，由于随机标签在各个时期内是固定且一致的，因此网络在开始后就开始拟合通过多次训练集。我们发现以下观察结果适合随机标签非常有趣：a）我们不需要改变学习率计划; b）一旦装配开始，它快速收敛; c）它完全融合到（过度）拟合训练集。还要注意“随机像素“和”高斯“开始比”随机标签“快收敛。这可能是因为随机像素，输入比最初的自然图像更加分离因此，属于同一类别，更容易为任意标签分配构建网络。

在CIFAR10数据集上，Alexnet和MLP都在训练集上收敛到零损失。阴影表1中的行显示了确切的数字和实验设置。我们还测试了随机标签ImageNet数据集。如附录中表2的最后三行所示，尽管它没有达到完美的100％前1精度，95.20％的准确度仍然是一百万随机的惊人1000个类别的标签。请注意，切换时我们没有进行任何超参数调整

从真正的标签到随机标签。通过对超参数进行一些修改，可以在随机标签上实现完美的准确性。网络也达到了~90％即使启用了明确的正则化器，前1精度也是如此。

**部分损坏的标签** 我们进一步检查神经网络训练的行为，在CIFAR10上从0（无腐败）到1（完全随机标签）的标签损坏程度不同数据集。网络适​​合所有情况的腐败训练集。图1b显示了随着标签噪声水平的增加，收敛时间减慢。图1c描绘了测试收敛后的错误。由于训练误差始终为零，因此测试误差与测试误差相同泛化错误。随着噪声水平接近1，泛化误差收敛到90％ - 随机猜测CIFAR10的表现

## 2.2实现

根据我们的随机化实验，我们讨论了我们的研究结果如何对推理泛化的几种传统方法提出挑战。

**Rademacher的复杂性和VC维度** Rademacher的复杂性是常用的灵活的复杂性度量假设类。 经验Rademacher的复杂性数据集上的假设类H {x1 ,.。。 ，xn}定义为

其中σ1,. 。 。 ，σn∈{±1}是i.i.d.均匀随机变量。这个定义非常类似我们的随机化测试。具体地，R n（H）测量H拟合随机±1二元标记的能力分配。虽然我们考虑多类问题，但考虑相关二元是很简单的相同实验观察所持有的分类问题。自从我们随机化

测试表明，许多神经网络完全适合训练集与随机标签，我们期望对于模型类H，R n（H）≈1。当然，这是一个平凡的上界Rademacher的复杂性不会导致实际设置中有用的泛化界限。类似的推理适用于VC维度及其连续模拟脂肪破碎维度，除非我们进一步限制网络。虽然巴特利特（1998）证明了对粉末破碎的约束根据网络权重的L1范数界限，该范围不适用我们在这里考虑的ReLU网络。这个结果被Neyshabur推广到其他规范等。 （2015），但即使这些似乎也没有解释我们观察到的泛化行为。

**统一稳定性** 抛开假设类的复杂性度量，我们可以考虑用于训练的算法的属性。 这通常与一些人一起完成稳定性的概念，如统一的稳定性（Bousquet＆Elisseeff，2002）。 一致的稳定性算法A测量算法对替换单个示例的敏感程度。 但是，它只是算法的一个属性，它没有考虑数据的细节或标签的分发。 可以定义较弱的稳定性概念（Mukherjee等，2002; Poggio等，2004; Shalev-Shwartz等，2010）。 最弱的稳定性措施是直接的相当于边界泛化误差并确实考虑了数据。 但是，它有很难有效地利用这种较弱的稳定性概念

# 3正则化的作用

我们的大多数随机化测试是在关闭显式正则化的情况下进行的。Regularizers当参数多于数据点时，它是理论和实践中减少过度拟合的标准工具（Vapnik 1998）。 基本思想是虽然原始假设太大而不能很好地概括，正规则有助于将学习限制在假设空间的一个子集中具有可管理的复杂性。 通过添加一个明确的正则化器，比如惩罚正则最佳解决方案，可行解决方案的有效Rademacher复杂性是显着的降低。

表1：CIFAR10数据集上各种模型的训练和测试准确度（百分比）。比较具有和不具有数据增加和重量衰减的性能。 结果还包括拟合随机标签。

正如我们将要看到的，在深度学习中，明确的正规化似乎扮演了一个相当不同的角色。 作为附录中表2的底行显示，即使有丢失和重量衰减，InceptionV3也是如此

如果不是很完美的话，仍能很好地适应随机训练组。 虽然没有明确说明，在CIFAR10上，初始和MLP仍然完美地适合随机训练集与重量衰减打开。 然而，打开重量衰减的AlexNet无法收敛随机标签。 至研究正则化在深度学习中的作用，我们明确地比较了深网的行为学习有和没有正规学者。

我们不是对深度学习引入的各种正则化技术进行全面调查，而是采用几种常用的网络架构，并比较行为时的行为。关闭装备正规的人。 涵盖以下正规制定者：

•数据增强：通过特定于域的转换来增强训练集。 对于图像数据，常用的变换包括随机裁剪，亮度的随机扰动，饱和度，色调和对比度。

•重量衰减：相当于权重的2正则化; 也等于很难约束欧几里德球的权重，半径由数量决定体重衰减。

•Dropout（Srivastava et al。，2014）：随机屏蔽掉图层输出的每个元素给定的辍学概率。 只有ImageNet的Inception V3在我们的版本中使用了dropout实验。

表1显示了CIFAR10上的Inception，Alexnet和MLP的结果，切换了数据的使用增强和减重。 两种正则化技术都有助于改进泛化

图2：隐式正则化因子对泛化性能的影响。 aug是数据增强，wd是重量衰减，BN是批量标准化。 阴影区域是累积的最佳测试精度，作为早期停止的潜在性能增益的指标。 （a）提早停止可以当其他正规制定者缺席时，可能会改善泛化。 （b）早期停止对CIFAR10不一定有帮助，但批量标准化可以稳定培训过程并改善概括。

性能，但即使关闭所有正规系统，所有模型仍然非常概括好。

附录中的表2显示了ImageNet数据集上的类似实验。 前18％的准确度当我们关闭所有正规化器时观察到下降。 具体来说，前1精度没有正则化为59.80％，而随机猜测仅在ImageNet上达到0.1％的前1准确度。更引人注目的是，随着数据增加，但其他明确的正规化器关闭，Inception能够实现前1精度达到72.95％。 实际上，似乎是使用增强数据的能力已知的对称性比仅仅调整重量衰减或防止低强度要强大得多训练错误。

在没有正规化的情况下，入门率达到80.38％，排名前5位，而报告的数量为ILSVRC 2012的获胜者（Krizhevsky等，2012）达到了83.6％。 因此，正规化是重要的是，通过简单地改变模型架构可以实现更大的收益。 这很难说正则化者被认为是一般化能力的基本相变深网。

## 3.1隐含的规则

早期停止被证明可以隐含地调整一些凸出的学习问题（Yao等，2007; Lin et al。，2016）。在附录的表2中，我们在括号中显示了最佳的测试精度沿着培训过程。它证实了早期停止可能会改善泛化性能。图2a显示了ImageNet上的培训和测试准确性。阴影area表示累积的最佳测试精度，作为潜在性能增益的参考提早停止。但是，在CIFAR10数据集上，我们没有观察到早期的任何潜在好处停止。

批量标准化（Ioffe＆Szegedy，2015）是一个规范化层响应的运算符在每个小批量内。它已被许多现代神经网络架构广泛采用例如Inception（Szegedy et al。，2016）和Residual Networks（He et al。，2016）。虽然没有明确地设计用于正则化，通常发现批量归一化可以改善泛化性能。 Inception架构使用了大量批量规范化层。测试批量规范化的影响，我们创建了一个“完全没有BatchNorm”的架构与图3中的Inception相同，只是删除了所有批量规范化图层。图2b比较了CIFAR10上两个Inception变体的学习曲线，并关闭了所有显式正则化器。规范化算子有助于稳定学习动态，但影响于泛化性能仅为3〜4％。精确的准确度也列在本节中表1中的“没有BatchNorm的开始”。总之，我们对显式和隐式正则化器的观察一直在暗示正规调整后，正规则可以帮助改善泛化性能。然而，正规化者不太可能像网络一样是推广的根本原因在所有正规人员撤职后继续表现良好。

# 4有限样本表达式

已经付出了很多努力来表征神经网络的表达性，例如，Cybenko（1989）;Mhaskar（1993）; Delalleau＆Bengio（2011）; Mhaskar＆Poggio（2016）; Eldan＆Shamir（2016）;Telgarsky（2016）; Cohen＆Shashua（2016）。几乎所有这些结果都处于“人口水平”显示整个域的哪些功能可以和不能由某些类表示具有相同参数数量的神经网络。例如，众所周知水平深度k通常比深度k-1更强大。我们认为在实践中更有意义的是神经网络在a上的表达能力有限样本的大小为n。可以将种群水平结果转移到有限的样本结果使用均匀收敛定理。但是，需要这种统一的收敛边界样本大小在输入的维度上是多项式的，在深度中是指数的网络，在实践中提出了一个明显不切实际的要求。

我们直接分析神经网络的有限样本表达性，并指出这大大简化了图像。具体而言，只要网络的参数p的数量是大于n，即使是简单的双层神经网络也可以表示输入样本的任何函数。我们说神经网络C可以表示d维中大小为n的样本的任何函数如果每个样本S⊆Rd与| S | = n和每个函数f：S→R，存在一个设置C的权重使得每个x∈S的C（x）= f（x）。

**定理1：**存在具有ReLU激活和2n + d权重的双层神经网络可以表示d维中尺寸为n的样本的任何函数。

证明在附录中的C部分给出，我们还讨论了如何实现宽度O（n / k）深度为k。 我们注意到，给出系数权重的界限是一个简单的练习矢量在我们的建设中。 引理1给出了矩阵A的最小特征值的界。这可以用于给出解决方案w的权重的合理界限。

# 5隐含规则：对线性模型的一种上诉

尽管由于许多原因深层神经网仍然是神秘的，但我们在本节中指出它并非如此必然易于理解线性模型的泛化源。 的确是有用的是吸引线性模型的简单情况，看看是否有可以帮助的并行见解我们更好地了解神经网络。

假设我们收集n个不同的数据点{（xi，yi）}，其中xi是d维特征向量和yi是标签。 让损失表示具有损失（y，y）= 0的非负损失函数，考虑到经验风险最小化（ERM）问题

如果d≥n，那么我们可以适合任何标签。 但是，有可能用如此丰富的模型进行推广

阶级没有明确的正规化？设X表示第i行为xT的n×d数据矩阵一世 。 如果X具有等级n，则系统为无论右侧是什么，方程式Xw = y都有无数个解。 我们可以找通过简单地求解该线性系统，ERM问题（2）中的全局最小值。但是，所有全球最小值是否同样普遍？ 有没有办法确定何时一个全局最小化将推广而另一个不会？ 理解最小质量的一种流行方法是解决方案中损失函数的曲率。 但在线性情况下，曲率所有最佳解决方案都是相同的（Choromanska等，2015）。 要看到这一点，请注意在何时yi是一个标量，

当y是向量值时，可以找到类似的公式。 特别是，Hessian不是一个功能选择w。 此外，Hessian在所有全局最优解决方案中都是退化的。如果曲率不能区分全局最小值，那么它是什么？ 一个有希望的方向是考虑主力算法，随机梯度下降（SGD），并检查SGD收敛的解决方案至。 由于SGD更新采用wt + 1 = wt - ηtetxit的形式，其中ηt是步长，et是预测误差损失。 如果w0 = 0，我们必须要求解具有w = Pn i =1αixi的形式

对于某些系数α。 因此，如果我们运行SGD，我们就知道w = XTα位于数据点。 如果我们也完美地插入标签，我们有Xw = y。 执行这两个身份，这减少到单个方程

它有一个独特的解决方案。请注意，此等式仅取决于之间的点积数据点xi。因此，我们得出了“核心技巧”（Scholkopf等，2001） - 尽管在一个¨中环形时尚。

因此，通过形成Gram矩阵（又称核矩阵），我们可以完美地拟合任何标签集。

对数据K = XX T并求解线性系统Kα= y得α。这是一个n×n线性系统当n小于十万时，可以在标准工作站上解决适用于像CIFAR10和MNIST这样的小型基准测试。

相当令人惊讶的是，拟合训练标签恰好可以​​为凸模型提供出色的性能。在没有预处理的MNIST上，通过简单地求解（3），我们能够实现1.2％的测试误差。请注意，这并不简单，因为内核矩阵需要30GB才能存储在内存中。尽管如此，该系统可在3分钟内在具有24个核心的商品工作站上解决带有传统LAPACK呼叫的256 GB RAM。首先应用Gabor小波变换数据再解决（3），MNIST的误差降至0.6％。令人惊讶的是，添加正规化不会改善任何一种型号的性能！CIFAR10的结果如下。简单地在像素上应用高斯核并使用否正则化达到46％的测试误差。通过随机卷积神经网络进行预处理使用32,000个随机过滤器，此测试错误降至17％误差2。进一步增加'2正则化将此数字减少到15％的误差。请注意，这没有任何数据扩充。

请注意，此内核解决方案在隐式正则化方面具有吸引人的解释。简单代数表明它等于Xw = y的最小`2范数解。那是，在所有完全符合数据的模型中，SGD通常会以最小的方式收敛到解决方案规范。构造Xw = y的解决方案非常容易，这些解决方案没有概括：例如，可以将高斯核拟合到数据中并将中心置于随机点。另一个简单的例子强制数据在测试数据上拟合随机标签。在这两种情况下，解决方案的规范是明显大于最小标准解决方案。

不幸的是，这种最小范数的概念不能预测泛化性能。对于例如，返回到MNIST示例，最小范数解的2范数没有预处理大约是220.通过小波预处理，规范跳跃到390.然而

测试误差下降了2倍。因此，尽管这种最小范数的直觉可能会提供一些指导对于新的算法设计，它只是概括故事的一小部分。

# 6结论

在这项工作中，我们提出了一个简单的实验框架，用于定义和理解一个概念机器学习模型的有效容量。 我们进行的实验强调了几个成功的神经网络架构的有效容量足以破坏训练数据。 因此，这些模型原则上足够丰富以记忆训练数据。这种情况对统计学习理论提出了一个概念上的挑战模型复杂性难以解释大型人工神经网络的泛化能力。

我们认为，我们尚未发现这些巨大模型的精确正式度量很简单。 我们的实验产生的另一个见解是优化仍在继续即使最终模型没有概括，也很容易实现。 这说明了原因

为什么优化在经验上是容易的，必须与泛化的真正原因不同。