## 通过影响力函数理解黑盒预测

1、KON P W, LIANG P. Understanding Black-box Predictions via Influence Functions[C]//ICML 17: 1885-1894.

2、BASU S, POPE P, FEIZI S. Influence Functions in Deep Learning Are Fragile[C]//ICLR 21.

## Understanding Black-box Predictions via Influence Functions

## (斯坦福大学)

# Importance of the application

* 机器学习系统经常被问到的一个关键问题是“系统为什么做出这个预测？”我们想要的模型不仅是高性能的，而且是可解释的。通过了解模型为什么会这样做，我们希望改进模型、发现新科学，并为终端用户提供影响他们行为的解释。
* 然而，许多领域中表现最好的模型——例如，用于图像和语音识别的深度神经网络——是复杂的黑盒模型，其预测似乎难以解释。解释这些黑盒模型的工作集中在理解固定模型如何导致特定预测，例如，通过在测试点周围局部拟合更简单的模型或通过扰动测试点以查看预测如何变化。这些工作根据模型解释了预测，但如何解释模型的来源？
* 模型参数最终来源于训练数据，通过跟踪模型的学习算法和训练数据来解决这个问题。为了形式化训练点对预测的影响，反事实地问：如果没有这个训练点，或者如果这个训练点的值发生了轻微的变化，会发生什么？

# Problems cannot be solved with existing methods

* 尽管影响力函数在统计方面有着悠久的历史，但它并未在机器学习中得到广泛使用。现有工作在模型不可微、非凸和高维的环境中的不适用性。

# Formal definition of the problem

考虑从某个输入空间（例如图像）到输出空间（例如标签）的预测问题。有训练点，其中。对于一个点和参数，令为损失，为经验风险。经验风险最小化由给出。假设经验风险在中是二次可微且严格凸的。

# What are the technical challenges with this approach?

* 如何计算和逆转经验风险的黑塞矩阵？对于个训练点，需要的操作，这对于具有数百万个参数的深度神经网络模型来说太费时了。
* 如何计算所有训练点的？

# Proposed Methods for these challenges

* 影响力函数通过训练数据的角度捕捉模型的核心思想。
* 增加训练点的权重：我们的目标是了解训练点对模型预测的影响。通过反事实询问来形式化这个目标：如果没有这个训练点，模型的预测会如何变化？
  + 首先研究由于从训练集中删除点导致的模型参数变化。然而，为每个移除的重新训练模型非常慢。
  + 影响力函数给了一个有效的近似值。这个想法是如果被小小地加权，则计算参数变化。
  + 本质上，形成了一个围绕的经验风险的二次近似，并使用了一步牛顿法。
  + 由于删除一个点与将其加权相同，因此可以线性地近似由于删除而导致的参数变化，而无需重新训练模型。
  + 应用链式法则来测量加权如何改变。特别是，增加对测试点损失的影响具有封闭形式的表达式：
* 扰动训练输入：通过研究不同的反事实来提出更细粒度的影响力概念：如果修改了训练输入，模型的预测将如何变化？
  + 虽然影响力函数似乎只适用于无穷小（连续的）扰动，但这种近似适用于任意：-加权方案允许在和之间进行平滑插值。
  + 通过在的方向上设置，可以构建的局部扰动，最大限度地增加处的损失。
* 高效计算影响力：
  + 技术挑战中的第一个问题在二阶优化中有很好的研究。它的想法是避免显式计算；相反，使用隐式黑塞-向量积（Hessian-vector products，HVPs）来有效地逼近，然后计算。
  + 这也解决了技术挑战中的第二个问题：对于每个感兴趣的测试点，可以预先计算，然后对每个训练点高效地计算。

**Experiments and Datasets**

* 实验
  + 影响力函数与留一法再训练

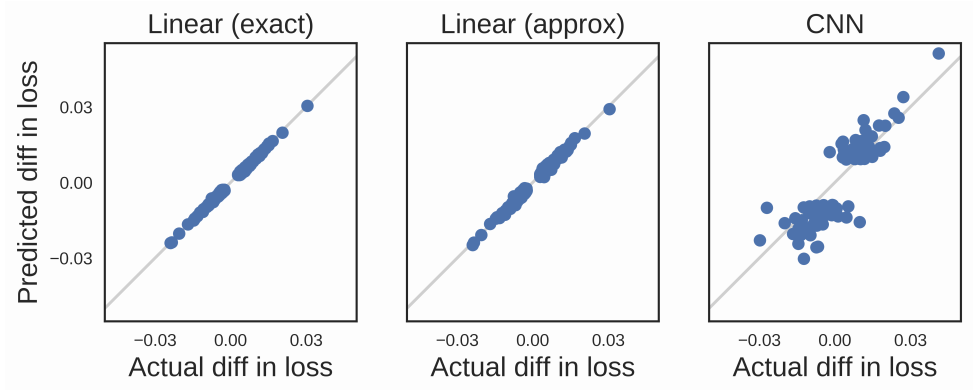


图1. 影响力函数与留一法再训练是一致的

为了研究使用影响力函数来近似去除训练点和重新训练的效果的准确性，将影响力函数与实际进行留一法重新训练进行了比较。使用十分类MNIST的逻辑回归模型，比较预测的和实际的变化。随机选择一个错误分类的测试点，它表现出的趋势是广泛的。

图1左图：对于影响力最大的500个训练点中的每一个，将影响力与移除该点并重新训练后测试损失的实际变化作图。HVP的逆是用CG计算的。

图1中图：相同的实验设置，但使用随机近似。

图1右图：CNN上相同的实验设置，使用CG对100个最具影响力的点进行计算。

* + 非凸性和非收敛性

如果通过提前停止或在非凸目标上使用SGD来获得参数，可能具有负特征值。对这类模型，影响力函数仍然给出有意义的结果。通过训练卷积神经网络进行500k次迭代来检查在非收敛、非凸设置中的表现。

图1右图：即使在这种困难的设置中，损失的预测变化和实际变化仍然高度相关。

* + 不可微分损失

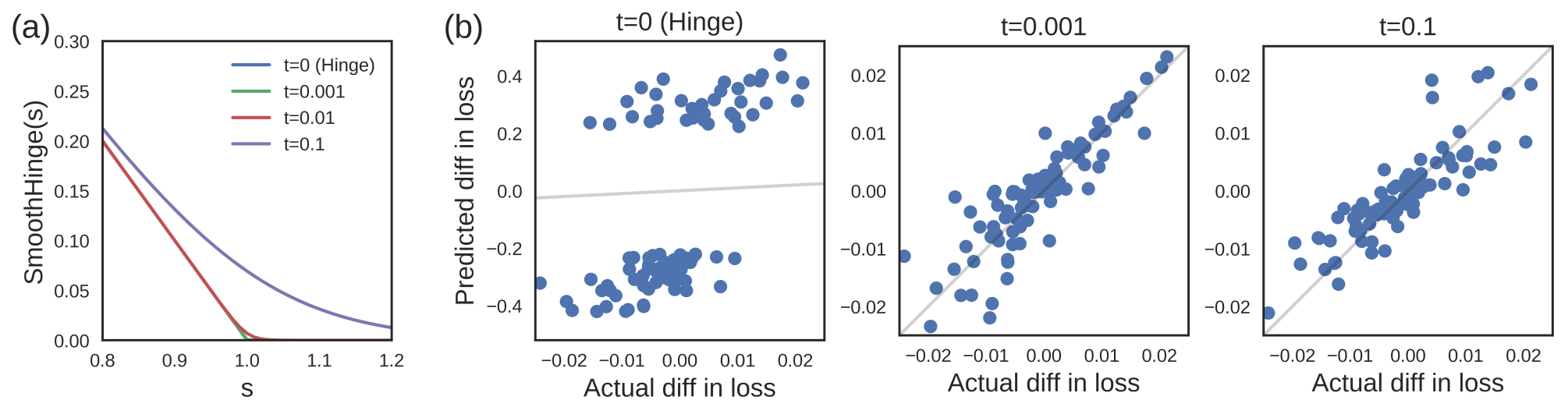


图2. 合页损失的平滑近似

当损失的导数不存在时会发生什么？在对不可微损失的平滑近似上计算的影响函数可以预测在留一法重新训练下原始不可微损失的行为。这种近似的鲁棒性表明，可以训练不可微分模型，并将不可微分组件换成平滑版本，以计算影响。在1和7的MNIST任务上训练了线性SVM。这个简单的分段线性函数类似ReLU，它导致神经网络中的不可微性。

图2（a）：为了计算影响力，用光滑合页来近似合页，它在时近似合页损失。通过改变，可以以任意精度近似合页损失。

图2（b）：使用随机的错误分类的测试点，比较了对100个最有影响力的训练点进行留一再训练后损失的预测与实际差异，其他测试点也有类似的趋势。

图2（b）左图：正如所预料的那样，影响函数无法准确预测变化：二阶导数不包含有关支持向量与合页的接近程度的信息，因此高估了的影响。

图2（b）中图：使用与以前相同的SVM权重，使用光滑合页计算影响力与原始合页重新训练导致的实际变化一致。

图2（b）右图：在很宽的范围内相关性仍然很高，当t太大时相关性会降低。

* + 理解模型行为

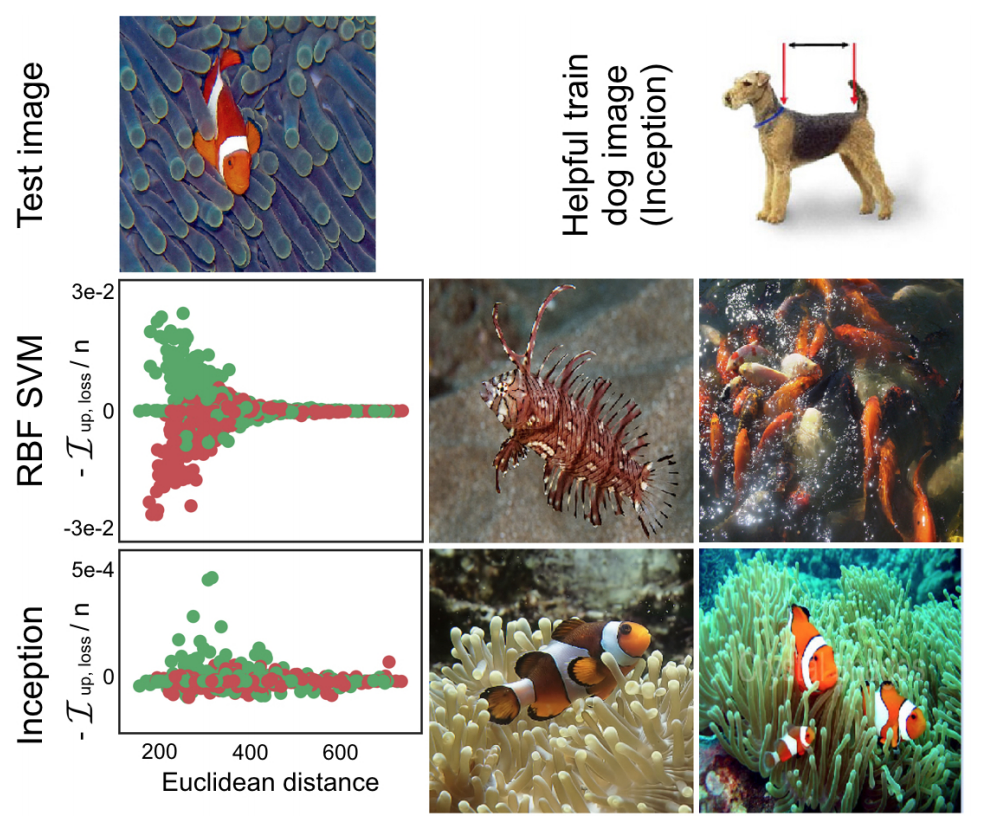


图3. Inception与RBF SVM

通过对给定预测负责的训练点，影响力函数揭示了模型如何依赖训练数据并从中推断，展示了两个模型以非常不同的方式做出相同的正确预测。比较了最先进的Inception v3网络和在ImageNet中提取的狗和鱼的图像分类数据集上带有RBF核的支持向量机。

图3上图：选择一个两个模型都正确的测试图像，并使用光滑合页来计算对SVM的影响。

图3下图：支持向量机的损失随原始像素距离的变化呈反比，像素空间中远离测试图像的训练图像几乎没有影响。Inception的影响与像素空间中的距离的相关性要小得多。在测试集中，每个模型展示两个最有帮助的训练图像，可以看到Inception网络发现了小丑鱼的特征，而支持向量机只是模式匹配训练图像。在支持向量机中，靠近测试图像的鱼（绿色点）多为有益的，而狗（红色点）多为有害的。相比之下，在Inception网络中，鱼和狗可能有助于或有害地将测试图像正确分类为鱼；一些最有帮助的训练图像是狗，对模型来说，它们看起来与鱼非常不同。

* + 对抗训练样本

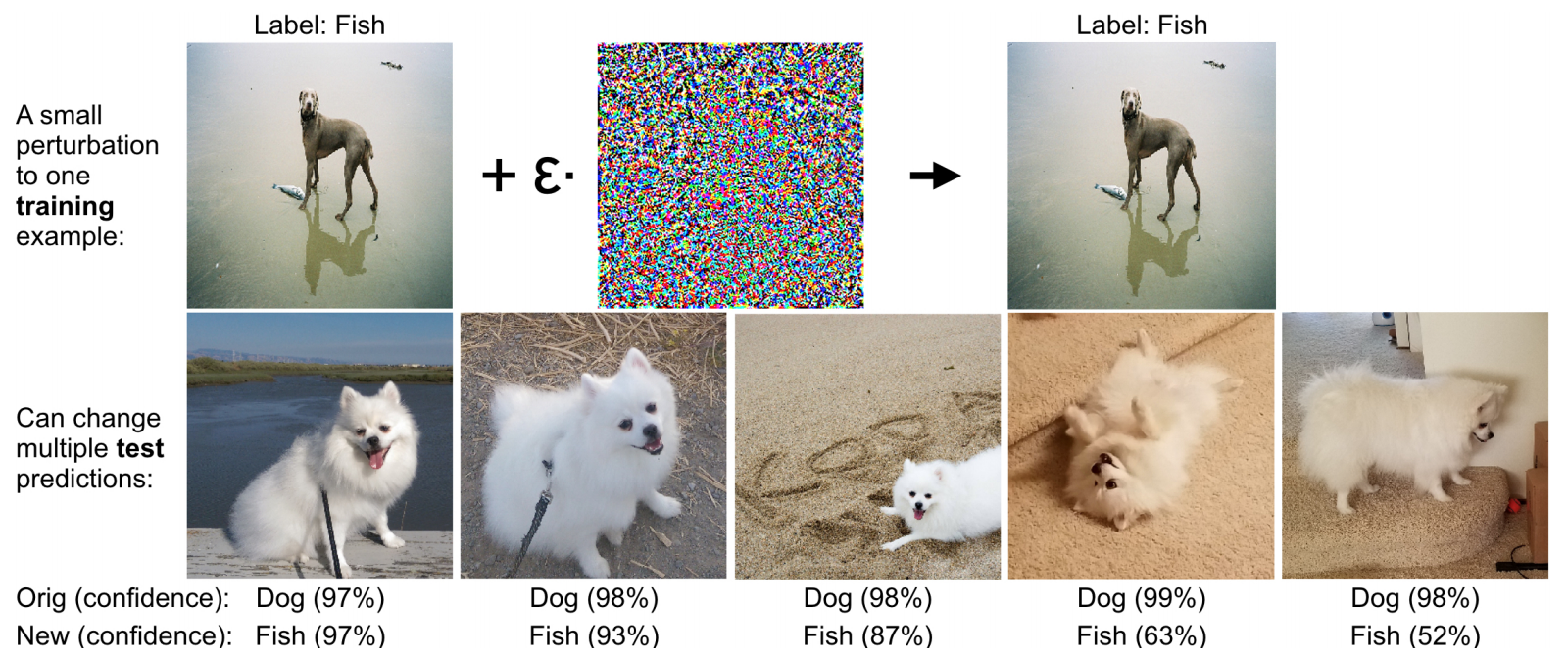


图4. 对训练集的攻击

攻击者可能会影响训练数据，对少量点施加很大影响的模型很容易受到训练输入扰动的影响，从而造成严重的安全风险。影响力函数可以用于制作视觉上不可区分的对抗性训练图像，并可以在单独的测试图像上翻转模型的预测。其核心思想是修改训练点以最大限度地增加损失。具体来说，对于目标测试图像，可以构造一个对抗版本的训练图像。

图4：在Inception网络上对狗和鱼进行了训练攻击测试。分别攻击每个测试图像，即使用不同的训练集攻击每个测试图像。通过增加平均损失来同时攻击多个测试图像，并发现单个训练图像扰动也可以同时翻转多个测试预测。首先，虽然像素值的变化很小，但在最终的Inception特征层中的变化明显很大：在像素空间中使用L2距离，训练值的变化小于训练点到类质心的平均距离的1%。这就有可能使攻击在视觉上难以察觉，但可以通过检查特征空间来检测。其次，攻击试图在低方差方向扰动训练样本，导致模型在该方向过拟合，从而错误地分类测试图像。第三，模糊或错误标记的训练图像是有效的攻击点：模型的置信度低，因此损失大，使它们具有很大的影响力。

通过度量影响力和损失的大小，模型开发人员可以量化模型在训练集攻击下的脆弱程度。

* + 修复错误标记的样本

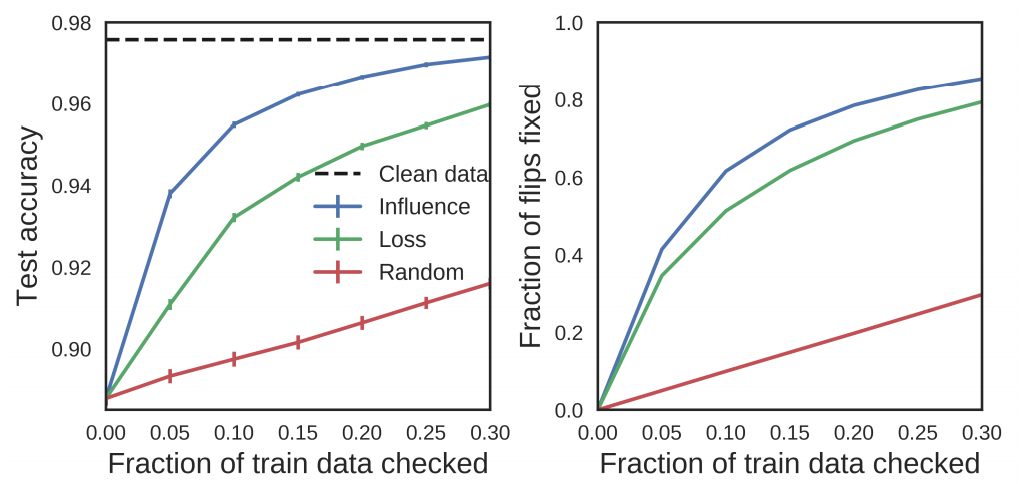


图5. 修复错误标记的样本

现实世界中的标签往往是有噪声的。即使人类专家能够识别出错误标记的样本，手动审查所有的训练数据也不可能的。展示了影响力函数可以帮助人类专家优先检查真正重要的样本。

关键思想是标记对模型影响力最大的训练点。案例研究是电子邮件垃圾邮件分类，它依赖于用户提供的标签，也容易受到对抗攻击。随机翻转10%的训练数据的标签，然后模拟手动检查一部分训练点，如果它们被翻转了，就纠正它们。使用影响力函数对训练点进行优先级排序以进行检查，能够在不检查太多点的情况下修复数据集。

图5：显示了测试精度（左）和检测到的翻转数据的比例（右）如何随检测到的训练数据的比例而变化，使用不同的算法来挑选要检查的点。

# Differences of this Method compared to others

影响力函数衡量局部变化的效果：当一个训练点的权重提高无穷小时，会发生什么？这种局部性允许推导出有效的封闭形式估计。

**Limitation of this work or your discovery**

* 影响力函数依赖于模型而不改变太多，如何理解模型的全局变化是一个开放问题，例如，该医院的患者群如何影响模型？

## Influence Functions in Deep Learning Are Fragile

## (马里兰大学)

# Importance of the application

* 在机器学习中，当训练样本的经验分布受到无穷小扰动时，可以使用影响力函数估计模型参数的变化。与重新训练模型以获取准确参数变化的过程相比，这种近似计算成本更低。因此，可以使用影响力函数来了解在测试时去除单个训练点（或一组训练样本）对模型预测的影响。利用损失函数的一阶泰勒近似，使用损失函数的梯度和黑塞矩阵计算的（一阶）影响力函数可以有用地解释机器学习模型，修复错误标记的训练样本，并对数据投毒攻击。
* 影响力函数通常在逻辑回归等模型上定义和研究，其中损失函数是凸的。然而，当损失函数的凸性假设被违背时，比如在深度学习中，影响函数的行为并没有被很好地理解，仍然是一个开放的研究领域。随着计算机视觉、自然语言处理、高风险应用如医学等领域的最新进展，解释深度模型预测变得尤为重要。这使得理解深度学习背景下的影响力函数至关重要。

# Problems cannot be solved with existing methods

* 现有工作认为影响力函数适用于非凸的深度网络。一个重要的问题是哪些类别的深度网络影响了函数的评估工作？各种深度网络的影响力估计是相当脆弱的。

# What are the technical challenges with this approach?

* 对于深度网络，有几个因素可能会对影响力估计产生影响：
  + 由于损失函数的非凸性，不同的扰动模型初始化会导致模型参数显著不同（损失值近似）；
  + 即使模型的初始化是固定的，网络的曲率值（即黑塞矩阵的特征值）在最优模型参数在深度模型中可能是非常大的，导致损失函数的泰勒近似误差，从而导致差的影响力估计；
  + 对于大型神经网络而言，计算影响力估计所需的精确黑塞向量积的逆在计算上可能非常耗时。因此，需要使用近似的黑塞向量积的逆技术，这可能是有错误的；导致低质量的影响力估计；
  + 在最优模型参数附近，不同的模型结构可能会有不同的损失形状，导致不同的影响力估计。

# Possible methods for this challenges

* 网络深度和宽度对影响力估计有很强的影响。当网络较浅时，影响力估计是相当准确的，而对于较深的模型，影响力估计往往是错误的。随着深度的增加，网络的曲率值不断增加。
* 权值衰减正则化对于在某些架构和数据集中获得高质量的影响力估计非常重要。
* 黑塞矩阵向量积的逆的近似技术是有误差的，特别当网络比较深时。这可能导致深度模型中的影响力估计质量较低。
* 在不同的数据集和架构中，测试点的选择对影响力评估的质量有重大影响。
* 在大规模的数据集中，由于模型的训练和收敛，即使是真实值的影响力估计(通过留一法重新训练)也可能不准确且有噪声。

**Experiments and Datasets**

* 数据集

首先在小的Iris数据集中研究影响力函数的表现，可以计算出精确的黑塞矩阵。

逐步增加模型和数据集的复杂性：使用小型MNIST来评估深度为6的小型CNN架构中影响力函数的准确性。

在标准MNIST和CIFAR-10数据集上训练的深度架构的影响力函数。

使用ImageNet来计算影响力估计。

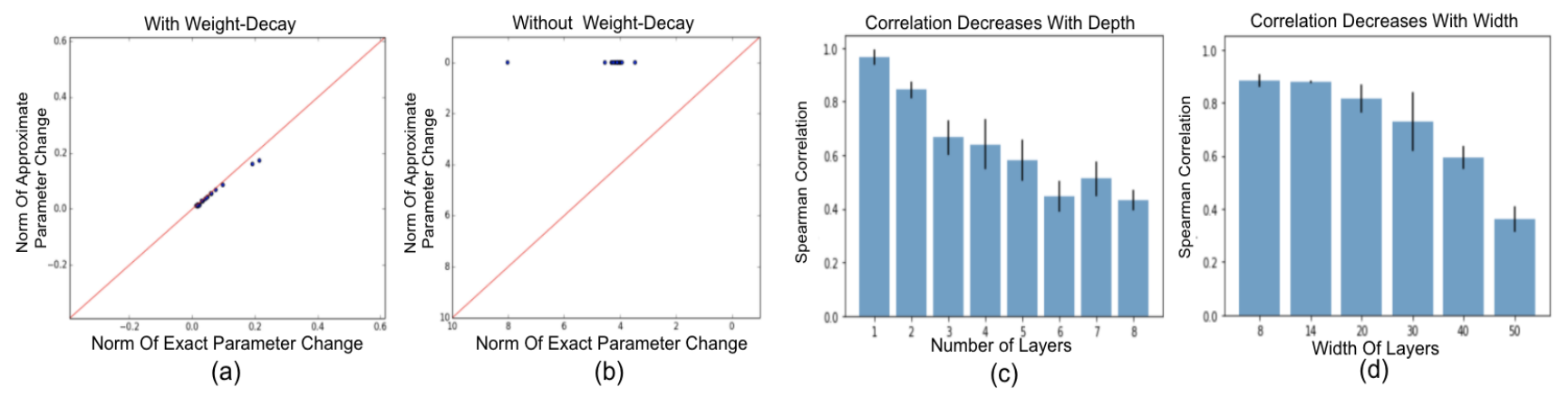


图6. Iris数据集的实验结果

一阶影响力函数假设当训练数据的经验分布受到无穷小扰动时，损失函数是凸的，模型参数的变化很小。然而，在非凸损失函数的情况下，这一假设一般是不正确的。根据经验，对于深度网络，泰勒差距受共同超参数的影响。

图6：使用影响力函数计算的参数变化范数与重新训练的比较。对于未经权重衰减正则化训练的网络，泰勒差距很大，导致低质量的影响力估计。

图7：当网络深度和宽度相当大（即过参数化）时，泰勒差距增加，大大降低了影响力估计的质量。

图7（a）：在Iris上，这种近似对影响力估计只有很小的影响。网络结构、超参数和损失的曲率是影响力估计的重要因素。

图7（b）：泰勒差距的增加在最优模型参数下评估的损失函数的曲率值有很强的相关性。

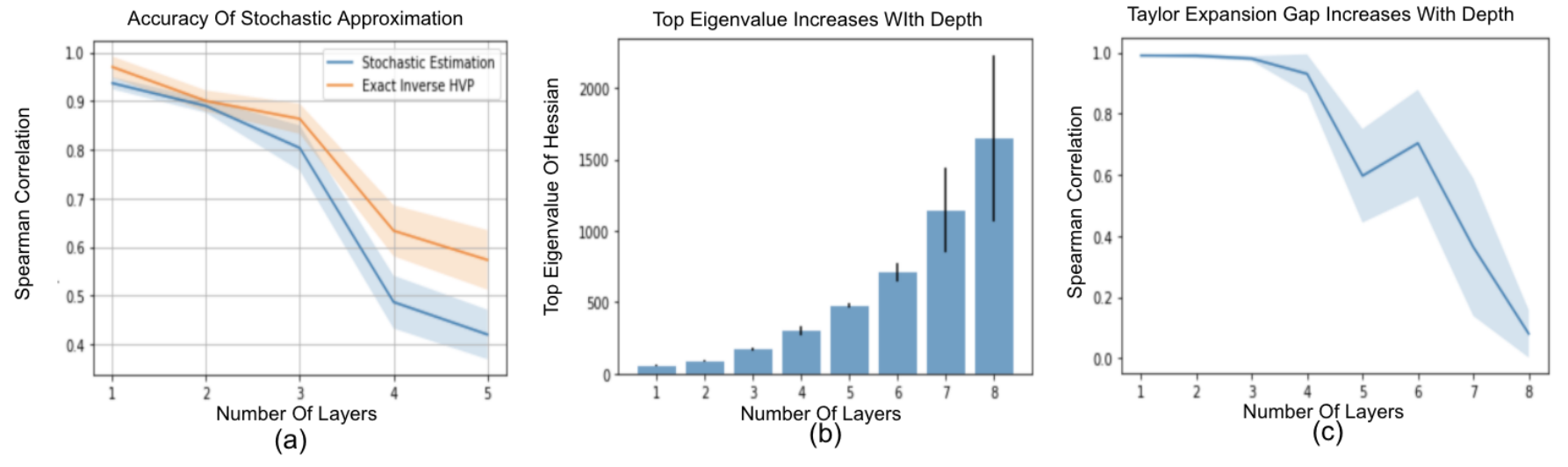


图7. Iris数据集的实验结果

* 权重衰减的影响

用于训练神经网络的一种常见的正则化技术是权重衰减正则化。在训练过程中，对模型参数的范数进行惩罚。训练了一个具有和不具有权重衰减正则化的简单前馈网络。对于使用权重衰减训练的网络，观察到影响力估计和真实估计之间的斯皮尔曼相关性为0.97。相比之下，对于未经权重衰减正则化训练的网络，斯皮尔曼相关估计下降到0.508。在这种情况下，黑塞矩阵是奇异的，因此在黑塞矩阵中加入阻尼因子0.001，使其可逆。为了进一步理解影响力估计质量下降的原因，在所有训练中比较以下度量：

a）通过再训练计算出的模型参数变化的范数

b）利用影响力函数计算出的模型参数变化的范数。

当网络在没有权重衰减的情况下训练时，用影响力函数计算的模型参数的变化与用重新训练计算的有很大的偏差。当模型没有权重衰减训练时，使用一阶影响力估计的泰勒展开的差距很大。

* 网络深度的影响

图6（c）：网络深度对影响力估计的质量有显著的影响。当网络的深度增加到8时，斯皮尔曼相关估计的显著下降。为了进一步理解当网络更深时影响力估计质量的下降，计算真实参数变化（通过重新训练计算）和近似参数变化（使用影响力函数计算）之间的近似差距。为了量化误差差距，计算最具影响力的16.6%样本中真实和近似参数变化范数之间的斯皮尔曼相关估计。随着深度的增加，真实参数和近似参数的范数变化之间的斯皮尔曼相关估计减小。

图7（b）：随着网络变得更深，损失函数的曲率会持续增加。网络的曲率信息可以作为真实参数与用影响力函数计算的近似误差差距的上界。

图7（c）：当网络深度大于5时，近似误差差距特别大。

* 网络宽度的影响

为了观察网络宽度对影响力估计质量的影响，通过逐步增加网络宽度来评估恒定深度的前馈网络的影响力估计。

图6（d）：随着网络宽度的增加，斯皮尔曼相关性不断下降。当网络宽度从8增加到50时，斯皮尔曼相关性从0.82下降到0.56。通过增加网络宽度来将其过参数化，会对影响力估计的质量产生强烈影响。

* 随机估计对黑塞矩阵向量积的逆的影响

对于大型深度网络，由于不能计算和求出准确的黑塞矩阵，使用随机估计计算黑塞矩阵向量积的逆。为了理解随机近似的有效性，分别用精确的黑塞估计和随机估计来计算影响力估计。

图7（a）：在不同的网络深度上，用随机估计计算的影响力估计比用精确的黑塞矩阵计算的影响估计有稍微低的斯皮尔曼相关性。网络越深，近似误差越大。

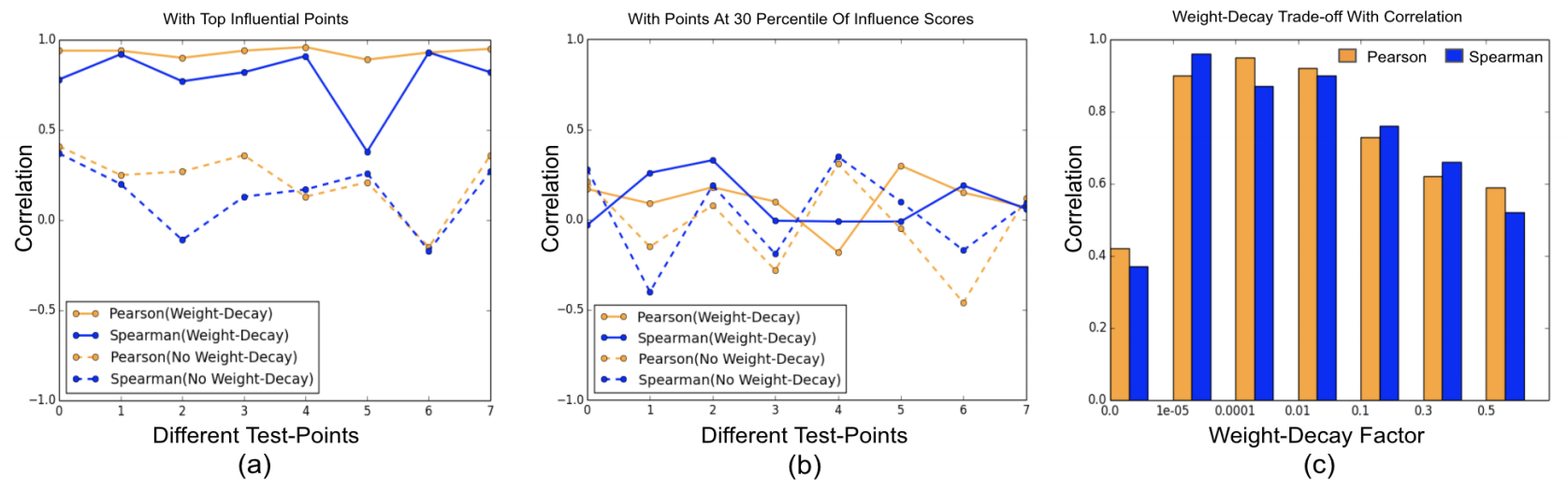


图8. 使用CNN架构的小型MNIST实验

图8（a）：当使用权重衰减进行训练并基于高影响力的点进行评估时，相关估计一致显著。

图8（b）：当使用位于影响力分数处于30%的训练样本集进行评估时，相关性估计显著降低。与真实值重新训练相比，仅对最具影响力点的影响力估计是准确的。

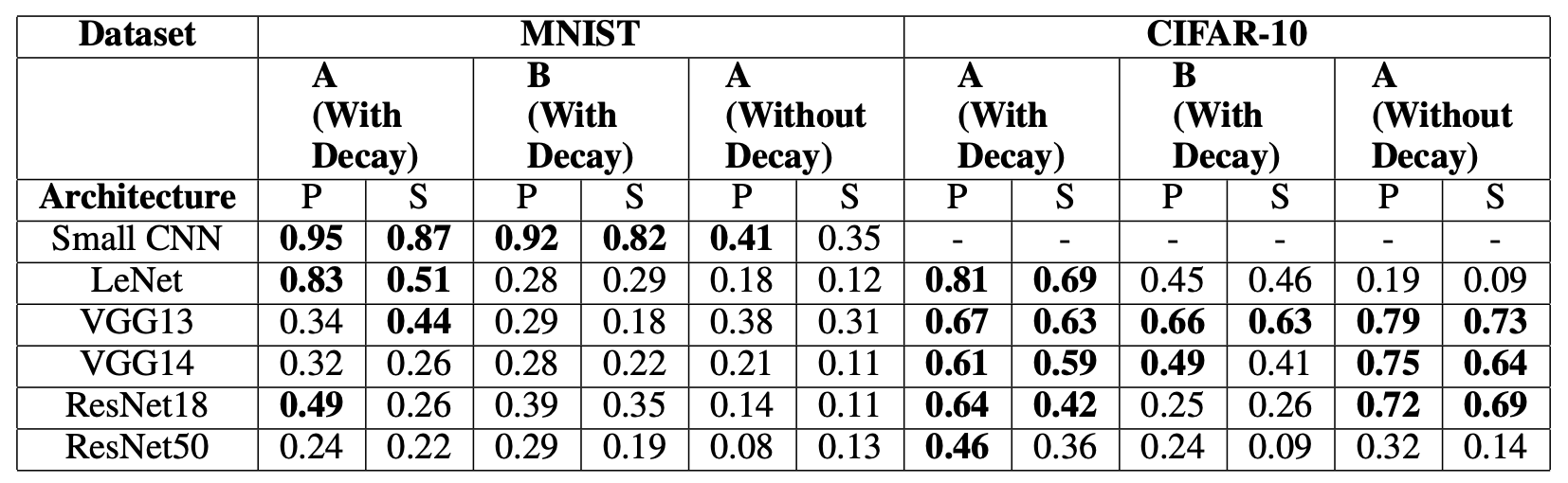
图8（a）和（b）：如果没有权重衰减正则化，两种情况下的影响力估计在所有测试点上都很差。测试点的选择对影响力估计的质量有很强的影响。例如，当使用权重衰减对网络进行训练，并且计算高影响力训练点的影响力估计时，在不同的测试点上，斯皮尔曼相关估计的范围从0.92到0.38，具有很高的方差。

为了进一步了解权重衰减对影响力估计的影响，使用不同的权重衰减正则化因子来训练网络。

图8（c）：权重衰减因子的选择对于获得高质量的影响力估计非常重要。对于特定的CNN架构，当权重衰减因子大于0.01时，相关性开始下降。

尽管影响力函数在非凸设置中有一些成功的应用，它的性能对实验的超参数以及训练程序非常敏感。

表1. MNIST和CIFAR-10的相关性估计



* 在MNIST上的结果

表1：对于损失最高的测试点，在小CNN和LeNet架构（用权重衰减正则化训练）中的影响力估计质量高。这些网络的参数分别为2.6k和44k，与实验设置中使用的其他网络相比，它们相对较小，深度也较小。随着网络深度的增加，影响力估计的质量持续下降。对于测试点损失在中位数的测试点，只有在较小的CNN架构中，影响力估计才有好的质量。

* 在CIFAR-10上的结果

对于CIFAR-10，在所有使用权重衰减正则化训练的架构中，损失最高的测试点的相关估计非常显著。对于大多数网络架构，相关估计高于0.6。然而，对于在损失的中位数评估的测试点，在大多数架构中相关性略微下降。在CIFAR-10上，即使没有经过权重衰减正则化训练的架构，在使用损失最大的测试点进行评估时，也具有高度显著的相关性估计。

在MNIST情况下，在浅层网络中，影响力估计是相当准确的，而在深度网络中，影响力估计的质量下降。对于CIFAR-10，虽然影响力估计是显著的，但在更深的网络，如ResNet-50，相关性略低。CIFAR-10影响力评估质量的提高可以归因于：对于类似的深度，在CIFAR-10上训练的架构比在MNIST上训练的架构更少地过参数化。过参数化降低了影响力估计的质量。

表1：测试点的选择对影响力估计的质量有相当大的影响。不同架构之间的影响力评估质量存在很大差异。对于小CNN和LeNet的影响力估计是相当准确的，而对于ResNet-50，MNIST和CIFAR-10的估计质量都下降了。

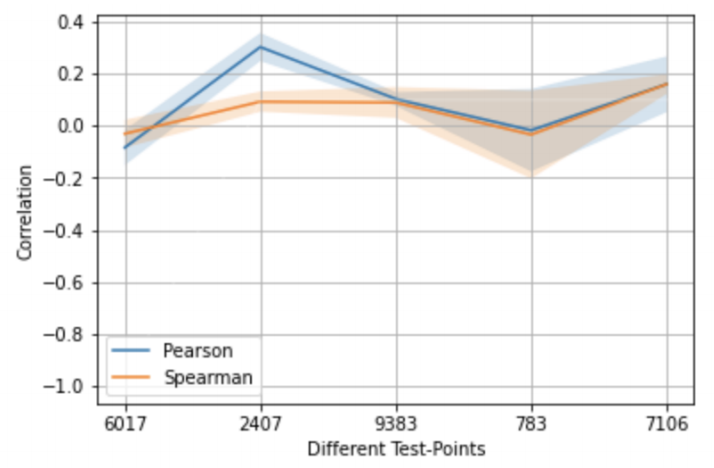
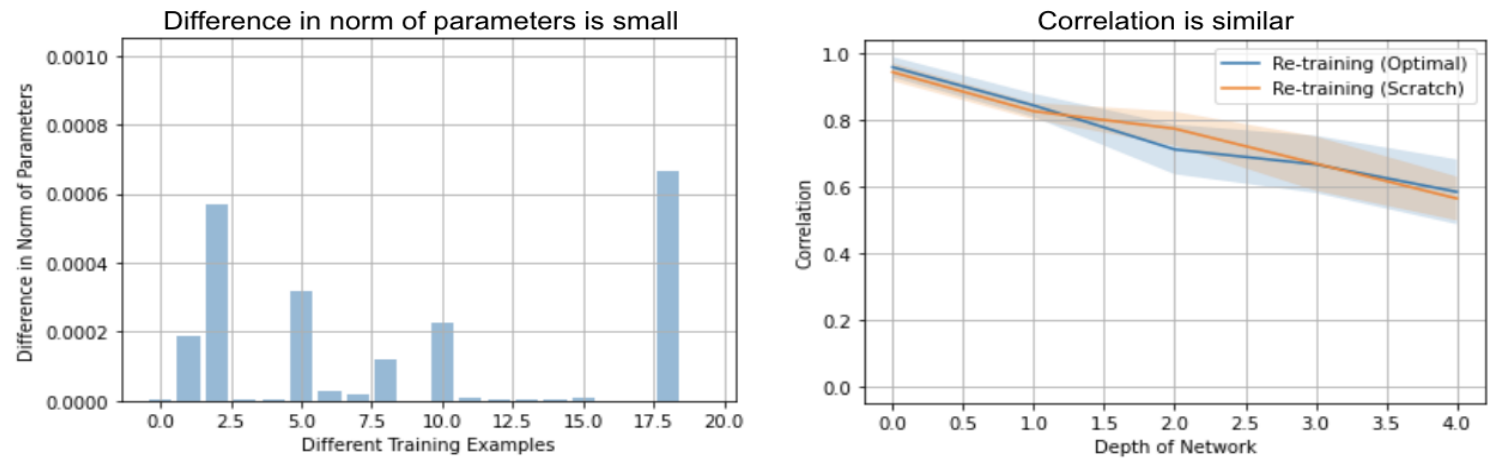


图9. CIFAR-100的影响

* 在CIFAR-100上的结果

在CIFAR-100的情况下，训练一个重衰减正则化因子为5e−4的ResNet-18模型。对多个模型初始化中损失最高的测试点和大约损失中位数的测试点计算影响估计。

图9：与MNIST和CIFAR-10的情况不同，相关估计的质量非常差。



通过重新训练模型获得的真实值估计之间的高可变性。这可能部分是由于原始模型没有完全收敛。

用所有训练点对原始模型进行额外2轮训练，并测量测试损失的变化。Top5测试的总体精度略有提高，达到92.336%(+0.034)，其中一个测试点的损失相对显著地下降了0.679。但是，另一个测试点小幅上升了0.066点。因此，损失值的变化可能会超过“留一法”重新训练程序的影响。

其次，计算权重梯度的二范数，在最优点附近应该接近于零，并将其与标准的预训练的ImageNet ResNet-50模型作为基线进行比较。这些范数分别为20.18和15.89，表明模型与基线具有相似的权重梯度范数。尽管考虑到有25.5M个参数，范数相对较小，但进一步对模型进行重新训练仍然会极大地改变一些样本的损失值，使真值估计具有噪声。

# Differences of this Method compared to others

在深度学习中提出一个全面分析成功和失败的影响力函数。证明了深度学习中的影响力函数通常是脆弱的。一些因素，如权重衰减、网络的深度和宽度、网络结构、随机近似和测试点的选择，都对影响估计的质量有很强的影响。在浅层架构中，如小CNN和LeNet，影响力估计是相当准确的，而在非常深和广泛的架构中，如ResNet-50，估计往往是错误的。在ImageNet规模中，影响力估计非常不精确。这些结果要求在深度学习的非凸情况下开发鲁棒的影响力估计。

**Limitation of this work or your discovery**

* 当前影响力函数在深度学习中的敏感性，需要开发用于大规模机器学习应用的鲁棒的影响力估计。