**Learning Deep ResNet Blocks Sequentially using Boosting Theory**

**摘要**：深神经网络由于反向传播的不稳定性而难以训练。具有单一循环的深层残差网络（ResNet）通过稳定梯度计算来缓解。我们证明了符合集成学习理论的ResNet体系结构。我们构建T个弱模块分类器，每个都包含两个T层，使得强学习器是一个ResNet。因此，我们引入了一种替代的Deep ResNet训练算法BoostResNet，它特别适用于不可微分的架构。我们提出的算法只需要一个简单的T“浅层网络”的连续训练。我们证明，如果我们训练的弱模块分类器比一些弱的基线更好，则训练误差以深度T指数地衰减。换句话说，我们提出一个弱学习条件，并在弱学习条件下证明ResNet的一个推论。我们的结果适用于一般的多类ResNets。证明了基于边际理论的泛化误差约束，并提出了ResNet在l1范数有界权重下抵抗过度拟合的网络。

**1介绍**

深层神经网络在机器学习方面取得突破性成就，特别是在图像分类和对象识别方面（Krizhevsky2012; Sermanet 2013; Simonyan＆Zisserman2014; Zeiler＆Fergus2014）。随着层次数的增加，非线性网络变得更加强大，从输入数据中获得更丰富的特征。实证研究表明，图像分类中的挑战性任务（He 2015; Ioffe＆Szegedy2015; Simonyan＆Zisserman2014; Szegedy2015）和对象识别（Girshick2015; Girshick 2014 ; He 2014; Long 2015; Ren2015）通常需要由数十或数百层组成的“深”网络。与浅网络相比，**理论分析**进一步证明了深度网络（Mhaskar＆Poggio2016）的能力。

然而，尽管深层神经网络具有内在的生存能力，但仍难以训练。反向传播的随机梯度下降（BP）（LeCun1989）及其变体通常用于解决非凸优化问题。存在对浅层和深层网络进行训练的主要挑战是梯度消失/爆炸（Bengio1994; Glorot＆Bengio2010）。近期工作提出了规范化技术（Glorot＆Bengio2010; LeCun2012; Ioffe＆Szegedy2015; Saxe 2013），以有效缓解问题。然而，在训练深度网络时，观察到令人惊讶的训练性能下降（He＆Sun2015; Srivastava 2015; He2016）：随着网络深度的增加，训练性能随着饱和点的增加而降低。这种训练性能下降是意想不到的，因为可以通过将深层网络的任何部分与浅网络相同，而其余的层作为同一映射来轻松构建与浅网络相同的深层网络。何（He2016）提出了一个残差网络（ResNet）学习框架来简化更深层的网络训练。并且通过添加同一变换，它们通过参考层输入将这些层明确地重新形成学习残差函数。在（Hardt＆Ma2016）中显示，identity loop可以缓解浅层网络中虚假局部最优的问题。 Srivastava等（Srivastava2015）介绍了一种新颖的架构，通过使用学习的**门控机制**来调节信息流，实现了几乎任意深度的网络优化。

虽然经验证据表明，这些深层残差网络比非残差网络更容易优化，但缺乏对这一观察结果的**理论依据**。例如对于这种深层残差架构的网络，训练或测试误差没有任何性能保证。此外，整个训练仍然依赖于在深层网络中易受非最佳解决方案影响的不稳定端对端反向传播（Ge et al2015）。

**1.1结果摘要**

我们提出了一个新颖的框架，多通道裂项求和提升（在第4节中定义），以表征第3节中的前馈ResNet。我们显示ResNet的顶级（最终）输出是其对连续的模块差异。提供了理论分析，如训练误差保证和裂项求和的泛化误差范围。

为了弥补随着深度增加而引起的误差退化问题，我们引入了一种称为BoostResNet的有保证的学习算法，对ResNet的模块进行了序列化。 BoostResNet自适应选择训练样本或更改成本函数（第4节定理4.2）。 BoostResNet可以保证ResNet的安全性：训练误差随着网络的深度呈指数级衰减。如第4.4节稍后所述，分析了BoostResNet的泛化误差，并提供了避免过度拟合的建议。我们的过程顺序地对每个残差块进行训练，只需要每个残差块提供对数据集标签的优于随机猜测的预测。

尽管有理论上的保证，我们的BoostResNet算法在端对端反向传播（e2eBP）训练惯例方面具有多个优势。首先，BoostResNet比e2eBP具有更高的内存效率，因为前者仅需要网络中的两个模块位于GPU中，而后者不可避免地将所有模块都保留在GPU中。优势对于深入网络的高效培训至关重要。此外，BoostResNet比e2eBP具有更高的计算效率，因为每个e2eBP步骤涉及通过整个深层网络的反向传播。

实验上，我们比较了BoostResNet和e2eBP两种类型的前馈ResNets，多层感知器残差网络（MLP-ResNet）和卷积神经网络残差网络（CNN-ResNet）在多个数据集上的比较。 BoostResNet在MLP-ResNet架构下显示出显着的性能提升。在CNN-ResNet下，观察到BoostResNet的收敛稍微更快一些。 BoostResNet在记忆和计算效率上的巨大优势也在实验中得到证明。我们的多通道裂项求和增强学习框架不限于ResNet，可以扩展到其他甚至不可微分的非线性假设单位，如决策树和张量分解。

**1.2相关工作**

训练深层神经网络在过去几年中一直是积极的研究领域。主要的优化挑战在于损失函数的高度非凸性质。解决这个优化问题有两个主要的方法，一个是选择具有更好几何属性的损失函数和网络架构，另一个是改善网络的学习过程。

**损失函数和体系结构选择** 在神经网络优化中，有许多常用的预定损耗函数和标准。例如，均方误差，负对数似然，边际准则等。关于选择或修改损失函数以预防诸如爆发/消失的梯度或缓慢学习的经验困难，有广泛的作品（Girshick，2015; Rubinstein＆Kroese，2013; Tygert等，2015）（Balduzzi等，2017）。但是，一般来说，没有严格的选择损失函数的原则。其他作品通过添加身份跳过连接（He et al，2016），考虑了MLP或CNN的变化，允许信息绕过特定层。然而，尽管经验性的成功取得了突破，但没有提供关于培训错误的理论保证。 Hardt等人（Hardt＆Ma，2016）已经显示出具有理论的线性神经网络中的identity loop的优点，然而线性设置在实践中是不切实际的。

**学习算法设计** 对BP改进有广泛的研究（LeCun等，1989），例如，动量（Qian，1999），Nesterov加速梯度（Nesterov，1983），Adagrad（Duchi等，2011）及其扩展Adadelta（Zeiler，2012）。最近，适应力矩估计（Adam）（Kingma＆Ba，2014）是动量和阿格达拉的组合，在实践中取得了巨大的成功。所有这些方法都是SGD的修改，但是我们的方法只需要一个任意的oracle，这不一定需要是SGD求解器。 Bengio等人（Bengio et al，2006）介绍了单一隐层凸神经网络，并提出了梯度增强算法来学习线性分类器的权重，然而将其方法的一般化到具有多个隐层的深层网络中不可能。 Shalev-Shwartz（Shalev-Shwartz，2014）提出了一个selfieBoost，提高了单个网络的准确性。我们的算法是不同的，因为我们改为构造分类器的集合。 AdaBoost（Cortes et al，2016）也考虑了分类器的集合，但是它们需要顶层和所有其他较低层之间的连接，这不符合标准的Res​​Net架构。

2预备

残差神经网络（ResNet）由称为模块的堆叠实体组成。 每个模块由多层神经网络组成。 常用的模块包括MLP和CNN。

在本文中，我们考虑训练示例（x，y）={（x1，y1），（x2，y2）,…,（xm，ym）}∈D，其中D是数据的分布。 我们将使用S表示样本。

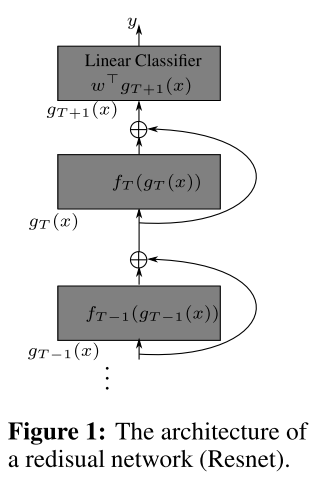
**ResNet的模块** 让每个模块将其输入映射到ft(·)，其中t表示模块的级别 ,每个模块ft（x）是具有n个通道的非线性单位，即ft（x）∈Rn。在多层次残差网络（MLP-ResNet），ft（x）是浅MLP，例如连接的非线性 ，其中Wt∈Rn×k，Vt∈Rk×n，σ是一个非线性算子，如S形函数或反函数。类似地，在卷积神经中网络（CNN-ResNet），功能ft（x）表示一个重新映射的网络（Resnet）介绍了第t个卷积模块。然后第t个模块输出gt + 1（x）



其中x是输入到ResNet的输入。请参见图1中ResNet，其中包含具有标识循环的堆叠模块。

**ResNet的输出** 由于等式（1）中规定的递归关系，第T个模块的输出等于低模数输出的总和，即 ，其中g1（x）= x。对于分类任务，ResNet的最终输出给出输入在表示gT + 1（x）之后，在线性分类器w∈Rn之后呈现x，





其中σ（·）表示从表示到标签的映射σ（z）：z→Y.例如，对于二进制分类，例如，或者多类分类。因此分类ResNet的输出的形式如下。深度ResNet的参数为{w，{ft（·），∀t∈T}}。 ResNet训练涉及训练样本（x1，y1），（x2，y2）.. ，（xm，ym）可用时对训练分类器w和模块ft（·）∀t∈[T]的权重进行训练。

**Boosting** Boosting（Freund＆Schapire，1995）假设有一种弱学习算法，在给定的训练示例中，产生一个弱分类器（a.k.a.基本分类器）ht（x）。提升的目标是提高弱学习算法的性能。提升背后的关键思想是选择弱分类器的训练集，以便强制它每次调用数据时推断新数据。最后学习算法将许多弱分类器ht（x）组合成单个组合强分类器 强分类器。

根据实证经验，ResNet解决了更深层神经网络中增加训练误差的问题。我们好奇的是，随着深度T的增加，ResNet的训练误差渐近收敛到0是否存在理论上的理由。更重要的是，我们有兴趣提出一种通过深层网络避免端到端反向传播（e2eBP）的新算法，因此不受非凸优化的不稳定性的影响。

**3 ResNet在裂项求和加成框架**

正如我们从方程式2回想的那样，ResNet确实具有与强化分类器相似的形式。关键的区别是，提升是一个估计假设的集合，而ResNet是估计特征表示F（x）=ΔT的集合。解决这个问题我们在每个ResNet模块ft（gt（x））的顶部引入一个辅助线性分类器wt，以构建一个假设模块。正式地，假设模块被定义为



在二进制分类设置中。因此其中 。在多类设置中，让C为类的数量，我们定义假设模块其中线性分类器Wt∈Rn×C是矩阵而不是向量。我们的分析适用于二分类和多类，但是我们将重点介绍二进制类，以便在主文中简单，并将多类分析推迟到附录E.

现在，假设模块的简单的集成不是ResNet，除非所有的辅助线性分类器wt都等同于ResNet的顶级分类器w。但是请注意，第t + 1个模块的输入值gt + 1（x）是第t个模块的输出ft（gt（x））+ gt（x），必须依次训练模块。因此，常见的辅助线性分类器假设阻止我们顺序地训练T假设模块，因此是不现实的。我们使用裂项求和和的想法设计一个弱模块分类器如下。

定义3.1。弱模块分类器定义为其中是假设模块。如果弱学习者被限制在弱模块分类器的形式，我们称之为增强框架的“裂项求和加成”。

**ResNet：弱分类器的集合** 回想一下，ResNet的第T个模块输出gT + 1（x），它被馈送到最终/最终线性分类器进行最终分类。 我们表明，弱模块分类器的集合相当于ResNet的最终输出。我们在引理3.2中正式声明它。

**引理3.2** 令第t个模块的输入值gt（x）为前一个模块的输出，即，gt + 1（x）= ft（gt（x））+ gt（x），则T弱模块的和分类器与深度为T，T的等式2中的输出F（x）相同



t = 1时弱模块分类器是，假设模块为有关证明见附录A. 请注意，我们会滥用方式来调用F（x）σ函数应用于F（x）之上，将输出映射到标签空间Y.我们现在分析第4节中的裂项求和式增援框架。

**4叠加求和加成二进制分类**

回想一下，弱模块分类器被定义为。我们限制边界分类器。在本文中，我们将假定和之间的协方差是非正的。我们提出一种学习算法，在下述弱学习条件下，训练误差能够随着模块数T的增加成指数型下降。

**4.1弱学习条件**

让 是假设空间的边界，其中是简单的权重。一个朴素的弱学习条件将是。但是这样的弱学习条件过于苛刻。即使当 接近于1，我们仍然会寻找那些表现比更好的弱分类器。所以，我们利用如下更弱的弱学习条件。

定义4.1（γ弱学习条件）。让.弱模块分类器满足关于一对分布（Dt，Dt-1）的γ弱学习条件，如果

弱学习条件是有理论推动并且在图5a所示的实验证明了。对每个弱分类器，被定义为表征训练样本之间真实标签和弱分类器h(x)相关性的“边”。4.1所展示的弱学习条件是非常容易满足的，因为只要比随机猜测稍好就行。假设空间的上界为1，我们选择

**4.2 BoostResNet**

我们现在提出了一种新颖的训练算法，利用叠加求和的方式进行二分类。特别地，我们在算法1和2中引入了一种深层ResNet的训练程序，称为BoostResNet，只需要顺序训练浅层。每个浅层残差结构配以线性分类的权重最后构成假设空间。 ResNet的权重在这些浅层ResNets上进行训练，辅助线性分类器最终可以不用（除了最后一层）。 因此，随着第t个模块的输出作为训练示例被馈送到下一个第t + 1个模块，训练算法是按照自下而上的逐个模块过程。

算法1

算法1 BoostResNet: 二分类的叠加求和方法

输入: m个带有标签的样本 其中 ，是一个阈值

输出: 和 ⊲ 无关参数，

1: 初始化 

2: 第0次迭代初始化样本权重: 

3: 当：

4:  

5: 计算 其中

6: 更新 其中

7: 

8: 循环结束

9: 

定理4.2。 [训练误差界限]使用算法1和2的T叠加和的boosting方法的训练误差随模块数T呈指数衰减



算法二2 BoostResNet: oracle 实现的残差模块

输入: 

输出: 

1: 

2: 

如果，弱模块分类器满足定义4.1中定义的γ弱学习条件，并且和之间的协方差是非正的

即使每个弱学习模块ht（x）只比随机（即γ> 0）稍好一些，Algorithm1和2的训练误差也能随着ResNet深度指数衰减而衰减。和之间的协方差是非正的，这表明弱模型分类器不应该是对抗性的，这对于ResNet是一个合理的假设。 有关多类分类的算法和理论保证，请参见附录E.

4.3 ResNet的Oracle实现在算法2中，oracle的实现相当于

（5）

实际上，有各种方法来实现方程（5）。例如，Janzamin等人。 （Janzamin等，2015）提出了一种张量分解技术，其分解通过与标签y组合的特征x的一些变换形成的张量，并且恢复具有保证的单隐层神经网络的权重。也可以使用反向传播，因为许多成果已经表明，基于梯度的训练在具有循环的浅网络上是相对稳定的（Hardt＆Ma，2016; He等人，2016）。

4.4本节分析了算法1的泛化误差，以了解其过拟合的情况。最终的强分类器或残差分类器为 现在我们定义样本的边界为

为了方便，我们考虑具有多个通道n的多层感知机残差网络，并且假定将层t上的神经元与其先前层神经元连接的权重向量是由L1范数约束的为 考虑最顶层的线性分类器w，使用L1范数界定，如 并且需要训练样本的无穷范数边界有限即 我们根据引理2（Cortes等，2016）导出引理4.3。

引理4.3。令 是关于的分布，S是从m个满足D分布的样本中随机选择的样例。对于，概率至少为 ，，强分类器 （ResNet）满足

(6)

其中 ，

从引理4.3，我们得到一个泛化误差的界限，称为上边界和网络复杂度。大的上界(对应大的)对于泛化精度有利而小的L1范数约束(对应小)能够有效控制网络复杂度，避免过拟合。网络复杂度的基础项是和网络深度成线性关系的。附录C有证明。

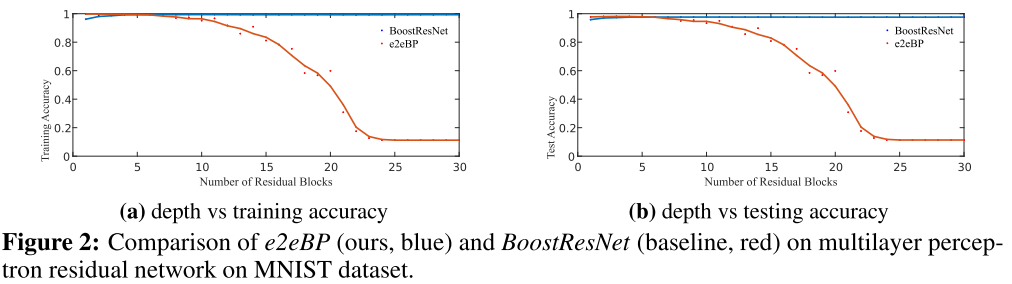
这个引理表明，预测精度更高和边界更大的较强弱分类器能产生更大的上界并减少过拟合。 越大，越小。对边界大的训练集并且，我们能够选择较大的让误差接近0.

**5实验**

我们比较了我们提出的BoostResNet算法与MNIST（LeCun等，1998）和街景房屋号码（SVHN）（Netzer等，2011）基准数据集的e2eBP训练ResNet。测试了两种不同类型的架构：多层感知器残差网络（MLP-ResNet）和卷积神经网络残差网络（CNN-ResNet）。在每个实验中，两种算法的架构是相同的，并且它们都被相同的随机种子初始化。我们的实验在Lua的Torch深度学习框架中编程，并在NVIDIA Tesla P100 GPU上执行。

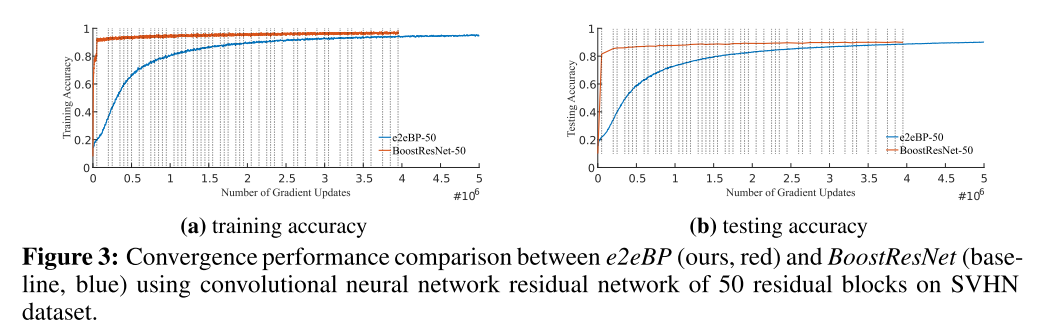
在我们的训练中，BoostResNet和e2eBP均使用小批量100。学习率以1e-2初始化，衰减速率为1e-4。 e2eBP中使用的优化方法是先前介绍的最新的自适应时刻估计（Adam）Kingma＆Ba（2014）变体。我们在BoostResNet中用于解决弱模块分类器的oracle也是Adam，但是可以扩展到张量方法Janzamin等。 （2015），决策树Safavian＆Landgrebe（1991）或其他用于不可微数据的非线性分类器。

**ResNet-MLP on MNIST**手写数字的MNIST数据库（LeCun等，1998）具有60,000个示例的训练集，以及10,000个示例的测试集。 数据包含10个类。 我们使用MNIST数据集测试MLP-ResNet上BoostResNet的性能，并将其与e2eBP基线进行比较。 每个残差块由具有单个1024维隐藏层的MLP组成。 BoostResNet和e2eBP之间的培训和测试错误在图2中作为深度的函数。令人惊讶的是，我们观察到e2eBP的训练误差降级，尽管ResNet的identity loop应该可以缓解这个问题。尽管存在identity loop，但是e2eBP最终都容易受到虚假的局部最优。我们提出的顺序训练程序BoostResNet缓解了梯度不稳定性问题，并且随着深度的增加而继续表现良好。

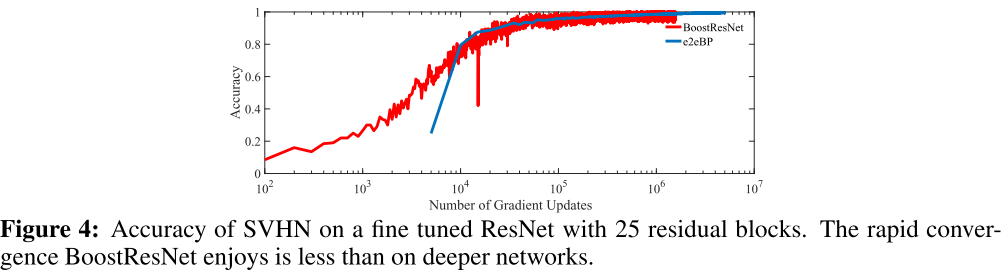


**ResNet-CNN on SVHN** SVHN（Netzer等，2011）是一个真实世界的图像数据集，从Google Street View图像中的房屋号获得，用于识别自然场景图像中的数字和数字，因此比MNIST更难。 许多图像在两侧包含一些垃圾箱。 数据集包含超过60万个数字图像，一个数量级更多的标签数据。 有604,388个数字的培训，26,032个数字的测试。 我们使用SVHN数据集测试CNN-ResNet上BoostResNet的性能，并将其与e2eBP进行比较。 每个残差

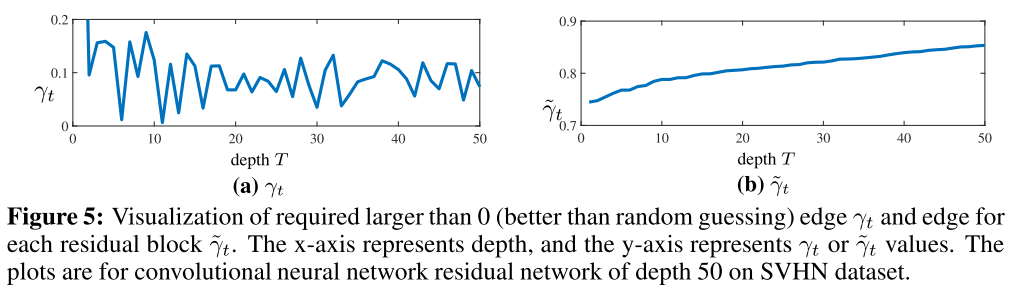
块由使用15个3×3滤波器的CNN组成。 BoostResNet和e2eBP之间的训练和测试错误如图3所示。在CNN-ResNet上，e2eBP的训练误差降级得到缓解，但与BoostResNet相比，学习速度相对较慢。



我们进一步研究e2eBP和BoostResNet在25个残差块和总共50层个较浅的微调网络上的性能。每个残差块包含卷积（5×5），批量归一化，ReLU，卷积（5×5），批量归一化， 和ReLU。 学习率固定为10e-4，无衰减。 精度结果如图4所示。尽管BoostResNet的参数没有被微调，但两种算法都是使用e2eBP微调网络架构和超参数进行比较的。



**弱学习条件检查** 图5中检查了受学习理论启发的弱点学习条件（定义4.1）。图5a中描绘了所需要的比随机猜测边γt更好，它始终大于0，因此我们的弱学习条件不成立，空虚的。 在图5b中，随着深度的增加，我们使用BoostResNet学习的表现越来越好（对于该分类任务）。



**计算和记忆效率** 值得注意的是，BoostResNet训练占用容量是高效的，因为训练过程只需要存储两个连续的残差块参数。 鉴于有限的GPU存储器是计算效率的主要瓶颈之一，由于降低通信开销和浅梯度转发和反向传播的速度，BoostResNet需要比深层网络中的e2eBP要少得多的训练时间。 对于一个线性分类器来说，记M1是所需要的，对于一个线性分类器来说，M2是需要的，由e2eBP的BoostResNet和M1T + M2的内存消耗是M1 + M2。 让翻转器需要在模块上进行梯度更新，一个线性分类器分别为C1和C2，BoostResNet和C1T + C2由BoostResNet计算成本为C1 + C2。实际上，50层CNN-ResNet需要3天时间才能使用e2eBP进行训练，但需要1天时间使用BoostResNet进行训练。

**6结论**

我们提出的BoostResNet算法在弱学习条件下实现了指数衰减（具有深度T）训练误差。与深层ResNet中的端到端反向传播相比，BoostResNet的计算效率要高得多。更重要的是，BoostResNet所需的内存与端对端反向传播相比是微不足道的。由于GPU内存有限，网络深度大，特别有利。我们的学习框架对于不可微数据是自然的。例如，我们的学习框架是可修改的，以采用弱度学习的方法使用张量分解技术（Janzamin et al，2015），将Tensor分解，具有理论保证的光谱学习框架应用于学习一层MLP。我们计划将我们的学习框架扩展到使用一般弱学习的不可微数据。