如何优雅地对深度神经网络进行训练

本文是对 Bengio 的论文《Understanding the difficulty of training deep deedforward neural networks》解读的第一部分。

主要的内容：

1、 对论文核心内容的解读，主要包含文中作者关于神经网络训练的观点
以及在实验中发现的“奇妙”的现象。

2、 关于深度学习，这里主要是对关于深度神经网络训练的难点进行剖析，并阐述了如何更好地进行深度神经网络训练。

这篇 Bengio 的论文很好地展示了神经网络参数的内在特性，人们发现文中很多具有指导性的思想并且从中得到很多训练神经网络的经验。目前非常火的深度学习框架 caffe、tensorflow 等等都有采用本文提到的关于权重、偏置初始化方法。仔细品读文章，我们会更好地理解：为什么对于随机初始化的深度神经网络，标准梯度下降优化算法会表现的如此拙劣，存在什么样的激活函数能够更好的工作等等。

Abstruct（摘要）

直到 2016 年以前，深度神经网络都是训练的非常失败，而对于新提出的训练方法，深度神经网络可以进行的很好。同时实验表明了深度网络比于浅层网络的优越性。深度神经网络得到成功的训练来自于：新的初始化方法或者说是新的训练机制。本文的主要目的是让大家更好的理解 **“为什么对于随机初始化的深度神经网络，标准梯度下降法表现的如此拙劣**。此外，最近相关领域内取得的成功以及来在未能能够帮助人们设计出更好的算法。”

我们首先注意到了非线形激活函数的影响。我们发现，s 型激活函数不太适合作为经随机初始化的深度神经网络的激活函数。**因为 s 型激活函数因其均值的原因（后文会有解释）能够让位于顶层的隐含层单元趋于饱和状态**。令人惊讶的是，我们发现处于饱和状态的神经元能够自己逃离这种饱和的状态，尽管转移这种饱和的情形发展的很慢，这也就解释了为什么我们能够看到神经网络在进行训练时，在很漫长的一个时间段里，曲线趋于平坦的

原因。而对于很少情况下使得神经元趋于饱和的其他非线形激活函数，往往表现的很好。

最后，我们知道：当与每层相关联的雅可比矩阵的奇异值远离 1 时，训练可能会变得更困难，我们研究激活值、梯度的值在神经网络层间训练是如何变化的。基于这些考虑，我们提出一个能够给神经网络训练带来更快的收敛的新的初始化方案。

1、 深度神经网络

深度学习方法旨在学习特征的层次结构，其具有由较低层特征的组合形成的较高层级的特征。它们包括用于广泛深度架构的学习方法，包括具有许多隐藏层的神经网络（Vincent 等人，2008）和具有许多隐层变量的图形模型（Hinton 等人，2006） Zhu 等人，2009; Weston 等人 2008）。自从受激发与于生物学和人类认知的理论吸引力和灵感，以及来自于在视觉和自然语言处理实验性的成功，他们得到了更多的关注。Bengio 评估和讨论的理论结果表明：为了能够学到一种可以表示高层抽象特征的复杂函数，我们需要深度神经网络。

大多数最近具有深度解构的实验结果从可以转变成深层监督神经网络的模型中得到，但初始化或训练计划不同于经典的前馈神经网络。为什么这些新算法比标准随机初始化和基于梯度的优化监督训练标准工作得更好？从最近对于无监督预训练的分析可以得出一部分答案，无监督预训练表明其充当正则化器，其在优化过程中，引人关注的是它的初始化的参数来自于在一个“更好的”分布中，对应于局部最小值，它是与更好的泛化能力相关联的。但是早期的工作已经表明：只有是贪婪层次程序的纯监督学习才可以给出更好的结果。**所以在这里，我们关注的不是无监督预训练或者是半监督学习准则给深层结构带来了什么，而是在多层神经网络中，什么东西会出错**。

我们的分析来自于监测层间及训练过程中激活值（观察隐含层单元的饱和情况）、梯度的调查实验。我们还得出对于不同激活函数（基于不同的激活函数会影响神经元饱和的想法）的选择以及初始化过程（因为无监督预训练是初始化的特例形式，而且它会有巨大的影响）带来的影响。

2、 实验设置及数据集（略）

3、 **激活函数以及训练时饱和带来的影响**

从激活函数的探索过程可以反映出如下两个方面是需要我们在训练神经网络中避免的：

一方面：激活函数的饱和性(这将使得梯度不能很好的进行反向传播)；

另一方面：过多的线性单元(它们不会计算一些令人感兴趣的的东西)。

3.1 Sigmod 函数实验

非线性激活函数 sigmod 的实验已经表明：由于自身非零均值的这个因素导致学习效率的下降，它能够引起海森矩阵出现奇异情况。在这一章节中，我们将看到在深度前馈神经网络中由 sigmod 函数引起的另一种“拙劣”表现。

我们将通过观察训练过程中激活函数值的变化过程来研究可能的饱和性，本节的图例中展现关于数据集 shapeset-3*2（见原论文）的结果，也可以从其他数据集中发现这些同样的结果。图 2 显示是深度结构网络中，每一个以 sigmod 为激活函数的隐含层在训练时， 其激活值的变化情况。层 1 是第一个隐含层的输出变化，这里一共有四个隐含层。图中给出了这些激活函数输出的均值和标准差分布情况。这些统计结果和直方图是在一个固定数量为 300 的训练集上通过观察训练的不同时间段的激活函数值的变化来统计得到的。

我们可以看到在起初阶段，最后一个隐含层的所有激活函数值都非常快地进入到他们较低饱和值 0。相反地，其他层的平均激活值在 0.5 以上，从输出层到输入层不断地在下降。我们发现，在以 sigmod 为激活函数的深度神经网络里，这种类型的饱和状态可以持续很长的时间。例如，五层网络从未逃离过这种状态。一个很惊讶的现象是对于具有中间数量的隐含层（这里是 4），则是可以脱离这种饱和状态。同时，处于顶层的隐含层可以“摆脱”这种饱和状态，第一个隐含层开始饱和且因此处于稳定的状态。

我们假设如下：神经网络的初始化是随机的，以及这样一个事实：以 sigmod 作为激活函数的饱和隐含层单元的输出为 0，基于以上两个因素共同导致的训练的表现。需要注意一点的是：以 sigmod 为激活函数且由无监督预训练（例如：RBMs）初始化的深度神经网络不会受到饱和行为的影响。我们提出的解释依赖于这样一个假设，那就是：随机初始化的神经网络的低层计算的变换对于分类任务来说是没有什么用的，它不像由无监督预训练获得的变换带来的结果。逻辑输出层$softmax(b + Wh)$起初更加依赖于 b（学

的非常快），而不是激活值 h，h 是源自于输入图像的(因为 h 将不以预测 y 的值而变化，而是可能跟 x 的其他或者更主要的变化相关)。因此，误差梯度函数可以通过将 h 趋近 0 得到 wh 趋于 0。类似于对称激活函数诸如 *tangent*、*softsign* 的这种情况，设置其值在 0 的周围是很好的，因为这样可以让梯度在反向传播的时候可以很好地流动（后面的文章会有分析）。然而，将 sigmod 函数的输出趋近 0（上面的函数在函数值为 0 时，其导数很大，而 sigmod 在值为 0 时，导数几乎为 0），将会让这些神经元进入饱和状态，这将会抑制梯度在反向传播中的流动，使得低层网络难以学到有用的特征。 最终，低层神经元会学习到到更有用的特征，高层神经元会逃离饱和的状态，但是整个过程是会比较慢的。需要主要的是，即使出现了这种情况，神经网络得到的这个训练结果是低质量的(也可以从泛化能力考虑)，这些是从对称激活函数中发现的，见图 11（图例是按原文安排的）。

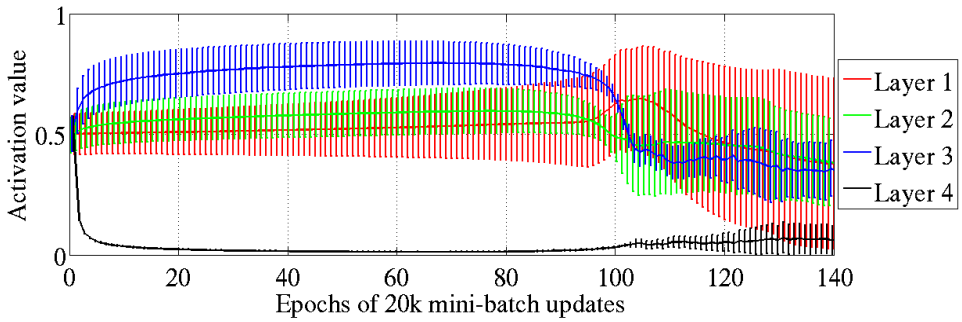

图 2：监督训练过程中处于不同层的激活值均值和标准差（y 轴）的分布情况。处于顶层的神经元很快处于饱和状态（值是 0）（整个学习过程一直在减速），在 100 次 epoch（迭代完训练集所有样本称为一次 epoch）后，逐渐脱离饱和状态。

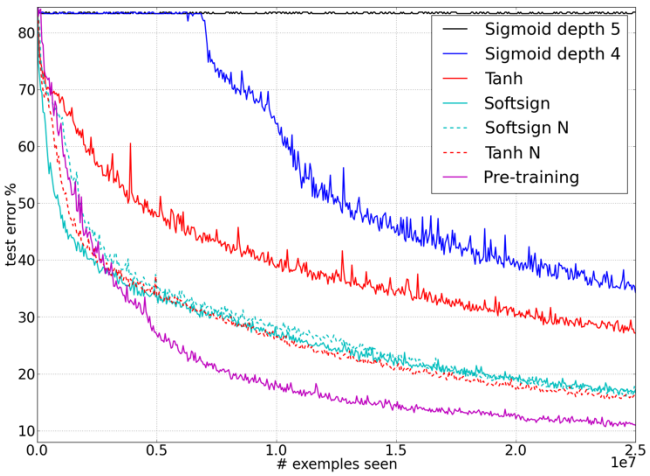

图 11:不同的激活函数及初始化方案在 Shapeset-3x2 数据集上在线训练误差曲线图（从上往下误差在下降）。激活函数值后面的 N 表示其使用标准初始化方案。

3.2 Hyperbolic tangent(双曲正切函数)实验

根据以上的讨论，由于双曲正弦函数是以 0 为对称的，以双曲正弦函数为激活函数的神经网络不会受到由 sigmod 网络中观察到的顶层隐含层神经元这种饱和状态行为的影响。然而，由于标准权重初始化服从$U\left[-\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right]$，我们可以从神经网络的第一层到最后一层，观察到一个连续出现的饱和现象，见图 3。为什么出现这种情况仍有待理解。

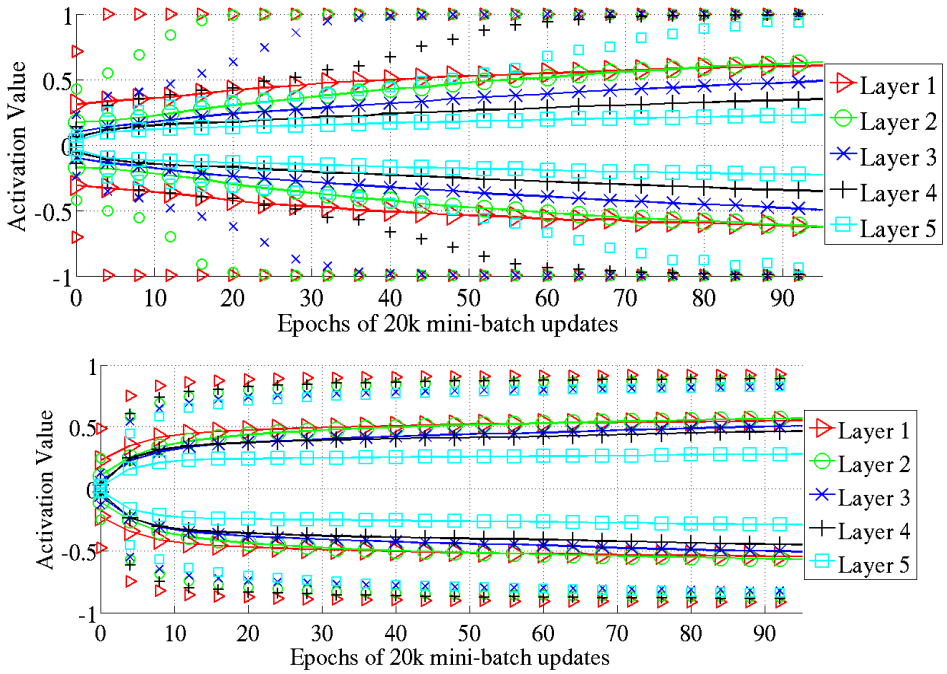

图3: 上图：98%（单独记录）的数据，来自于深度网络学习过程中双曲正切激活函数值的标准差（实心线标记）分布。我们看到第一个隐含层先进入饱和状态，其次是第二个隐含层，以此类推。下图：98%（单独记录）的数据，来自于深度网络学习过程中softmax函数值的标准差（实心线标记）分布。这里不同的层很少饱和并且不同层表现的一样（这很重要）。

3.3 Softsign 函数实验

softsign 函数和双曲正弦函数类似，但是对于饱和这种情形，考虑到其尾部的平滑特性（多项式形式而不是指数形式），其表现可能表现的不同。从图 3 中，对于双曲正弦函数，我们看到不会发生一层接着一层的饱和情况。在刚开始的时候发生的非常快，然后就很缓慢了，所有层的权重都一起朝着更大的方向学习。

同时，我们可以从训练的末期看到激活值的直方图跟激活函数是双曲正弦函数神经网络是不同的（图 4）。这里，后者激活函数分布产生的模型

在正负 1 的两端或者在 0 的附近，而激活函数是 softsign 神经网络的激活值的模型是分布在“其”周围（0 附近的线性状态和-1、1 平坦状态）。在这些区域，函数显示出显著的非线性，但是梯度可以很好的传播。

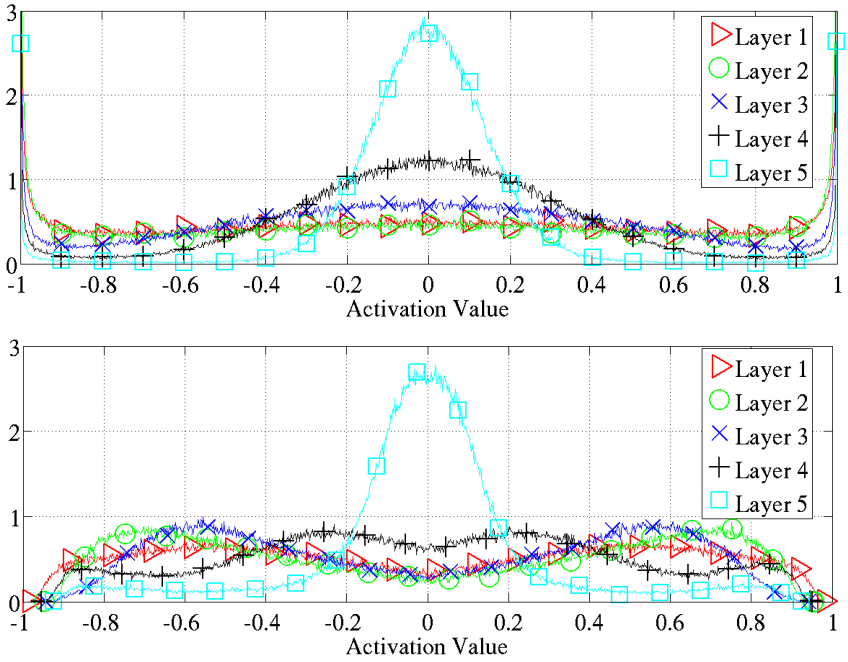

图 4:激活值直方图归一化。上图：双曲正切函数，可以看到低层饱和性。下图：softmax 函数，很多值分布在（-0.8，-0.6）和(0.6,0.8)周围，这里的单元并不会饱和且是非线性的。

4、 梯度学习及其传播

5、 误差曲线和结论

以上是对 Bengio 的论文《Understanding the difficulty of training deep deedforward neural networks》解读的第一部分,接下来的第二、三等等部分我们会逐步深入理解论文的主要内容。