

深度学习中激活函数权重初始值的选取

孙洲鑫* 王文松

SUN Zhou-xin WANG Wen-song

摘要

在深度学习的过程中, 激活函数的选取是非常重要的, 对于不同的激活函数如何选择合适的权重初始值是本文研究的重点。本文主要就激活函数权重初始值的选取提出问题, 根据相关资料的查询选择几个较为特殊的权重初始值在不同激活函数上进行验证实验, 得出不同激活函数对应的最佳的权重初始值的组合。

关键词

深度学习; 激活函数; 权重初始值; MNIST 数据集; sigmoid; ReLu; tanh

doi: 10.3969/j.issn.1672-9528.2020.09.032

0 引言

众所周知, 在最近几年里, 深度学习 (Deep Learning, DL)^[1] 的研究是最热的方向之一。深度学习是机器学习研究的一个分支, 在机器学习中, 参数的设定是通过大量事例进行类比得出, 深度学习则是通过人工神经网络 (Artificial Neural Networks, ANNs) 通过不同隐藏层^[2] 的学习进而得出最优的参数组合。

在深度学习的过程中, 重点就是激活函数^[3] 和权重^[4]

的选取, 合理的组合会使得学习速率加快, 同时可以降低损失函数的值。激活函数的选取一般都是根据实际的情况来做一定的改变。权重的参数是根据大量的自学习进而得到有效的参数组合, 对于不同的激活函数, 如何选择合适的权重初始值就成了本文所要研究的重点。

1 权重初始值的设定

在深度学习的过程中, 常用的激活函数主要有 sigmoid() 函数^[5] 如式 (1) tanh() 函数^[6] 如式 (2) 线性整流函数^[7] (Rectified Linear Unit, ReLU) 如式 (3) 等, 其数学表达式分别如下:

* 长江大学 湖北荆州 434023

参考文献:

- [1]KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Image net classification with deep convolution neural networks[C]. International Conference on Neural Information Processing Systems,2012:1097-1105.
- [2]DONG C,CHEN C L,HE K, et al. Learning a deep convolution network for image super-resolution[C].European Conference on Computer Vision,2014:184-199.
- [3]DONG C,LOY C C,HE K, et al. Image super-resolution using deep convolution networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,38(2):295-307.
- [4]DONG C,LOY C,YUBIN D et al. Compression Artifacts Reduction by a Deep Convolutional Network[C].Proceeding of the IEEE International Conference on Computer Vision,2015:576-584.

- [5]WENZHE S,JOSE CABALLERO,ROB B et al. Real-Time Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolution Neural Network[C].2016 IEEE Conference on Computer Vision and Pattern Recognition(CV PR),2016:1874-1883.
- [6]LEDIG C,THEIS L,HUSZAR F et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network[C].2017 IEEE Conference on Computer Vision and Pattern Recognition,2017:105-114.
- [7]Aharon M. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. TSP 2006,54(11):4311-4322.

(收稿日期: 2020-07-19 修回日期: 2020-08-11)

$$y = 1/(1 - e^{(-x)}) \quad (1)$$

$$y = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \quad (2)$$

$$y = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3)$$

对于不同的激活函数, 应该如何选择权重初始值是其中的重点。权重初始值如果选择过大, 那就会导致数据的过拟合^[8]或者降低学习的泛化能力^[9]; 权重如果选择 0 就会失去学习能力。

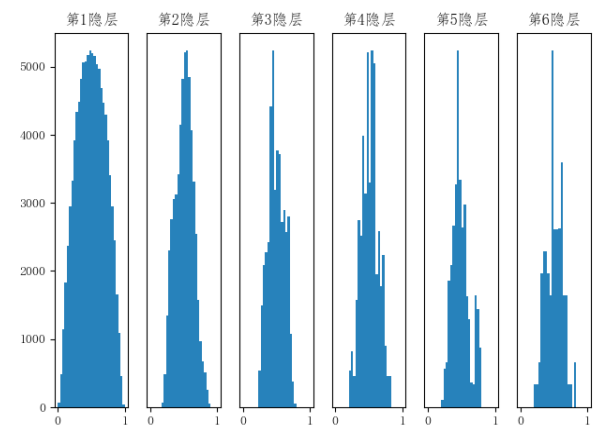
为了抑制过度拟合以及提高深度学习的泛化能力, 那就引出了权值衰减^[10] (weight decay) 的概念。权重衰减就是一种通过减小权重参数的值来达到深度学习的目的。如何正确减小权重的初始值才是本文研究的重点。权重初始值一般设定为: $0.01 * np.random.randn(10, 100)$, 即标准差为 0.01 的高斯分布。本文筛选以下四种权重初始值作为测试用例, 分别是标准差为 1, 0.01, $\sqrt{1/n}$, $\sqrt{2/n}$ 的高斯分布。来验证不同的激活函数与权重初始值的组合会产生怎么的效果。

2 权重初始值对不同激活函数的影响

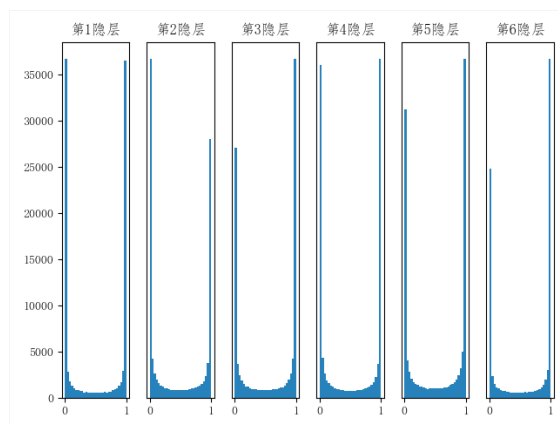
本次实验采用一个 6 层神经网络进行测试, 每层 100 个神经元, 然后利用高斯分布生成 1000 个数据作为输入的原始数据, 并传入神经网络中, 通过直方图绘制出各层激活值的数据分布进行分析, 得出对应激活函数的最佳权重初始化参数。

2.1 激活函数为 sigmoid() 的处理

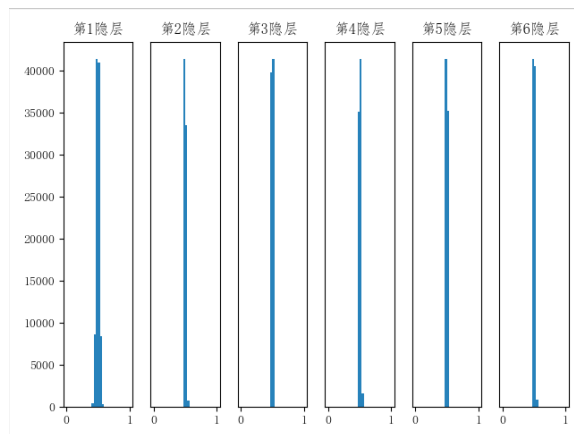
sigmoid 函数是一种常见的 S 型曲线函数, 在深度学习中被用作神经网络中的激活函数, 输出 (0, 1) 上的随机数。选择激活函数为 sigmoid() 后, 分别对四种权重初始值进行测试实验, 如下图 1 所示。



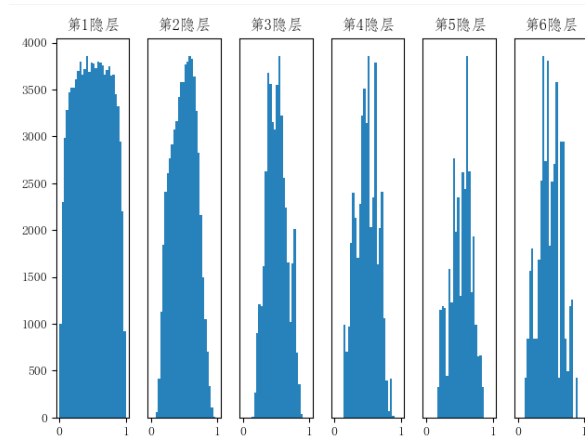
(a) 使用标准差为 1 的高斯分布的各层激活值



(b) 使用标准差为 0.01 的高斯分布的各层激活值



(c) 使用标准差为的高斯分布的各层激活值



(d) 使用标准差为的高斯分布的各层激活值

图 1 激活函数为 sigmoid()

由图 1 (a) 可以看出各层的激活值呈现出偏向 0 或 1 的分布, 当向 0 或 1 靠近时其导数就越接近于 0, 所以在反向传播中梯度的值不断减小, 最后消失, 并且随着隐藏层数的增多这种现象更严重。

由图 1 (b) 可以看出数据的分布在 0.5 的附近, 所以出现图 1 (a) 梯度消失的问题, 但是其输出值都是类似的, 说

明它的表现力不足,因为多个神经元的输出一致的原因,学习能力就丧失了。

由图 1(c) 可以看出数据分布的前期成一个“钟”型,很符合本文所说的概率模型,随着隐藏层的增加,其数据分布有所倾斜,但是也表现出了比之前更加宽的分布,这说明在各个层之间传递的过程中广度有所增加,所以其学习能力也相应有所提高。

图 1(d) 和图 1(c) 类似,但是从整体上看可以得出,其跳动行相比于图 1(c) 有点大,出错率相对较高,对于图 1(a)、图 1(b) 来说,其学习能力较强。

对比以上四个激活值的直方分布图可以看出,对于 sigmoid() 激活函数的最优的权重初始值为标准差是的高斯分布的值,由于其随着隐藏层数的增加,其广度有所增加,说明在不断地学习,通过神经网络的反向传播进一步进行参数的设置是实现深度学习的重要途径。

tanh 激活函数与 sigmoid 函数一致,区别在于其对称点不同,它们的最优权重初始值都是标准差为的高斯分布的值,随着层数的增加,图形变得越倾斜,同时呈现出范围更大的分布,随着数据过程中适当广度的增加,导致激活函数的表现力不受限制,可以实现高效的学习。

2.2 基于 MNIST 数据集及 ReLU 激活函数对初始权重的比较

该实验是以 MNIST 数据为原始输入数据,通过 5 层神经网络进行自学习。MNIST 数据集是由 Yann 提供的手写数字数据库文件,包括训练集 60000 张和测试集 10000 张手写字图像。通过不同的权重初始值,比较它们损失函数值的大小进而判断哪种初始值更加适合的测试。在 MNIST 数据集进行自学习时采用的激活函数为 ReLU 函数,每迭代 100 次取一个测试点,形成的折线图如下图 2 所示。

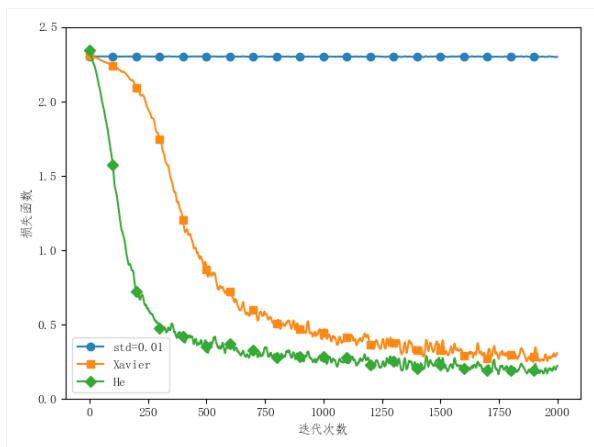


图 2 MNIST 数据集下不同初始值权重的损失函数值折线图

从图中可以很明显的看出,对 MNIST 数据集进行测试的时候,He 参数(标准差为高斯分布的值)的损失函数的值明显优于其他两个初始权重值。从图中结果可知, std=0.01 时

无法进行学习,与前文当激活函数为 ReLU 函数时的激活值分布图得出的结论一致,正向传递很小的值时,反向传播易造成梯度消失,从而丧失学习能力。当权重初始值为 Xavier (标准差为高斯分布的值)时,学习能力也相对较强,但对于和 He 参数相比, Xavier 参数还是有点慢,所以当激活函数为 ReLU 时,选择的权重初始值为 He 参数。

3 总结

综合之前的所有实验结果,可以得出,在神经网络的学习中,权重初始值的选取是非常重要的。如当激活函数为 sigmoid 时,应选择权重初始值为 Xavier 参数;当激活函数为 ReLU 时,应该选择权重初始值为 He 参数;当激活函数为 tanh 时,应该选择权重初始值为 Xavier 参数。只有合理的设置权重参数值才能更好地达到深度学习的效果。

参考文献:

- [1] 胡侯立, 魏维, 胡蒙娜. 深度学习算法的原理及应用 [J]. 信息技术, 2015(2):175-177.
- [2] 成凌飞, 贺扬, 张培玲等. 压缩深层神经网络隐藏层维度对其分类性能的影响 [J]. 电子科技, 2019(1):72-75.
- [3] 刘宇晴, 王天昊, 徐旭. 深度学习神经网络的新型自适应激活函数 [J]. 吉林大学学报:理学版, 2019,57(4):1-2.
- [4] 斋藤康毅. 深度学习入门基于 Python 的理论与实现 [M]. 陆宇杰, 译. 北京:人民邮电出版社, 2019:26-28.
- [5] Elfwing Stefan, Uchibe Eiji, Doya Kenji. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning [J]. NEURAL NETWORKS, 2018(107):3-11.
- [6] G. Ososkov, P. Goncharov. Shallow and deep learning for image classification [J]. Optical Memory and Neural Networks, 2017,26(4):221-248.
- [7] Wang, G, Giannakis, GB, Chen, J. Learning ReLU Networks on Linearly Separable Data: Algorithm, Optimality, and Generalization [J]. IEEE TRANSACTIONS ON SIGNAL PROCESSING, 2019,67(9):2357-2370.
- [8] 陶砾, 杨朔, 杨威. 深度学习的模型搭建及过拟合问题的研究 [J]. 计算机时代, 2018(2):14-17+21.
- [9] 刘威, 刘尚, 白润才, 等. 互学习神经网络训练方法研究 [J]. 计算机学报, 2017(6):1291-1308.
- [10] 索明何, 程乐. 基于 PLSR 的深度信念网输出权值确定方法 [J]. 控制工程, 2018(4):668-676.

(收稿日期: 2020-08-02 修回日期: 2020-08-24)