DOI:10.13873/J.1000-9787(2018)02-0050-03

ReLU 激活函数优化研究*

蒋昂波, 王维维

(浙江大学 超大规模集成电路设计研究所,浙江 杭州 310027)

摘 要: 门控循环单元(GRU)是一种改进型的长短期记忆模型(LSTM)结构,有效改善了 LSTM 训练耗时的缺点。在 GRU 的基础上,对激活函数 sigmoid,tanh,ReLU 等性能进行了比较和研究,详细分析了几类激活函数的优缺点,提出了一种新的激活函数双曲正切线性单元(TLU)。实验证明:新的激活函数既能显著地加快深度神经网络的训练速度,又有效降低训练误差。

关键词: 门控循环单元; 神经网络; 激活函数; 双曲正切线性单元

中图分类号: TP301.6; TN911 **文献标识码**: A **文章编号**: 1000-9787(2018)02-0050-03

Research on optimization of ReLU activation function*

JIANG Ang-bo, WANG Wei-wei

(Institute of Very Large Scale Integrated Circuit Design, Zhejiang University, Hangzhou 310027, China)

Abstract: Gated recurrent unit(GRU) is an improved long short term memory model(LSTM) architecture, it is effective to improve training time-consuming features of LSTM. Performance of some activation functions such as sigmoid tanh, rectified linear units(ReLU) are compared and researched on the basis of GRU architecture and analyze their advantages and disadvantages in detail. Propose a novel activation function named tanh linear unit (TLU). The experiment shows that the new activation function can not only speed up training speed of deep neural networks, but also effectively reduce training error.

Keywords: gated recurrent unit(GRU); neutral network; activation functions; tanh linear unit(TLU)

0 引 言

长短期记忆模型[1](long short term memory, LSTM)作为递归神经网络(recursive neural network, RNN)非常重要的一个改进,能够有效记忆和利用历史信息,已经在文本分析、语音识别、图像处理等众多领域得到了成功应用,极大促进了深度学习领域的发展。但其结构的复杂性导致训练模型的过程比较耗时。本文采用 Cho K 在 2014 年提出的门控循环单元[2](gated recurrent unit, GRU)结构,是一种在结构上改动比较大的 LSTM 变体,其将 LSTM 结构中的遗忘门(forget gate)和输入门(input gate)合并成一个更新门(update gate),使得深度神经网络在运算的候少了很多矩阵乘法,从而改善了 LSTM 训练耗时的缺点,在数据量很大的情况下,GRU 能节省更多的时间。

激活函数是 GRU 等深度神经网络结构的核心所在,目前常见的激活函数包括 sigmoid[3] 系的 sigmoid 和 tanh 函数,ReLU 系的 ReLU[4],LReLU 函数等。但 sigmoid 系的函数在后向传递的过程中出现了梯度消失[5](gradient vanishing)问题,极大地降低了训练速度。

ReLU 函数能够有效缓解梯度消失问题,其以监督的方式训练深度神经网络,无需依赖无监督的逐层预训练,显著提升了深度神经网络的性能。Krizhevsky A[6] 等人对常用的激活函数 ReLU, sigmoid 和 tanh 函数进行了测试,证明了 ReLU 函数的性能优于 sigmoid 系函数。

但 ReLU 也存在着致命的缺点。首先,ReLU 函数的输出大于 0,使得输出不是 0 均值,即均值偏移[7](bias shift),易导致后一层的神经元得到上一层输出的非 0 均值的信号作为输入,使得网络参数 W 计算困难。其次,随着训练的推进,部分输入会落入 ReLU 函数的硬饱和区,导致对应权重无法更新。均值偏移和神经元死亡共同影响了深度神经网络的收敛性和收敛速度。

本文在 GRU 结构上对 sigmoid 系的激活函数和 ReLU 系的激活函数进行了对比和研究,详细分析了两类激活函数存在的优缺点,并在此基础上设计了一种新的激活函数双曲正切线性单元(tanh linear unit, TLU),其综合了 sigmoid 系和 ReLU 系函数的优点,既能有效缓解梯度消失问题,也有效地避免了均值偏移现象。实验证明:这种新的函数在提

收稿日期:2017-01-09

* 基金项目:国家"863"计划资助项目(2012AA041701);浙江省教育厅资助项目(Y201010035)

升神经网络训练速度和降低误差率方面的作用非常显著。

1　激活函数的对比与研究

1.1　sigmoid 系激活函数

sigmoid 系函数包括 sigmoid 和 tanh。sigmoid 函数定义为

$$f(x)=\frac{1}{1+e^{-x}} \tag{1}$$

其函数图像如图 1 所示。

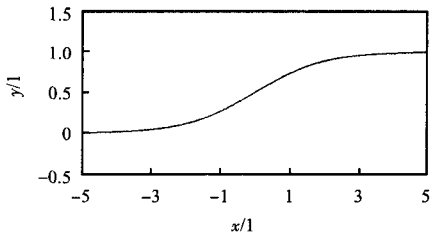

图 1　sigmoid 函数

从函数表达式和图像可见,sigmoid 函数具有软饱和性[8]:在定义域内处处可导,当输入非常大或非常小时,其图像的斜率趋近于 0,即导数逐渐趋近于 0。这种性质导致了梯度消失现象,使得深度神经网络一直难以得到有效训练,是阻碍神经网络发展的重要原因。

具体地,深度神经网络在使用梯度下降算法求解网络参数 W 时,在后向传递过程中,sigmoid 函数向下传导的梯度包含了一个自身关于输入的导数 $f'(x)$,当输入落入饱和区时,$f'(x)$ 的值趋近于 0,导致向底层网络传递的梯度变得非常小,使网络参数 W 很难得到有效训练。

sigmoid 函数也存在均值偏移的缺点,从函数图像可以看出,sigmoid 函数的值域为 $\{\forall x, y=f(x)\geqslant 0)\}$,则其输出均值必然非负,导致了 sigmoid 函数在训练一些超深网络时会出现训练结果不收敛的问题。

tanh 函数是 sigmoid 函数的一个变体,缓解了 sigmoid 函数所遇到的均值偏移问题,定义为

$$\tanh(x)=\frac{1-e^{-2x}}{1+e^{-2x}} \tag{2}$$

其图像如图 2 所示。

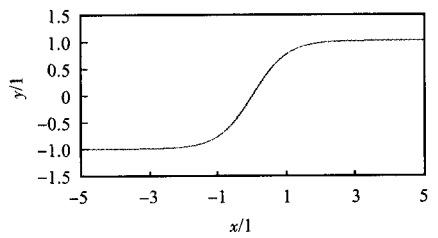

图 2　tanh 函数

从图像以及函数表达式中可以看出,tanh 函数也具有软饱和性,因此,也存在梯度消失的缺点。但其值域为 $[-1, 1]$,因此,输出均值趋近于 0,缓解了均值偏移问题,使得随机梯度下降(stochastic gradient descent,SGD)更接近自然梯度(natrual gradient),从而降低了计算网络参数 所需的迭代次数,提高了深度神经网络的训练速度。

1.2　ReLU 系激活函数

ReLU 函数有效解决了 sigmoid 系函数的梯度消失问题,但依然存在均值偏移的缺点。定义为

$$f(x)=\max(0,x) \tag{3}$$

其函数图像如图 3 所示。

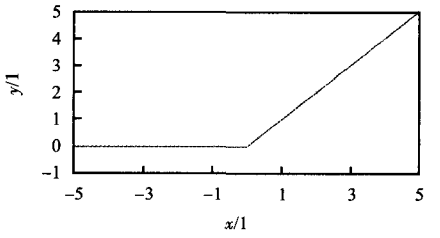

图 3　ReLU 函数

从函数表达式和图像可知,当 $x\geqslant 0$ 时,其导数为 1,因此,ReLU 函数能够在 $x\geqslant 0$ 时保持梯度不衰减,可以有效地缓解梯度消失问题。

当 $x<0$ 时硬饱和[8]。如果有输入落入此区域,则该神经元的梯度将永远为 0,不会再对任何数据有激活作用,即神经元死亡,直接导致计算结果不收敛。而且,ReLU 函数在 $x<0$ 时输出为 0,使得整体输出均值大于 0,无法缓解均值偏移问题。

PReLU 函数为 ReLU 函数的改进版本,具有非饱和性,能够缓解均值偏移问题和神经元死亡问题,其定义为

$$f(x)=\begin{cases}x, x>0\\ \alpha x, \text{其他}\end{cases} \tag{4}$$

其函数图像如图 4 所示。其中 $x<0$ 部分的图像根据其斜率 α 变化,一般 $\alpha=0.25$。

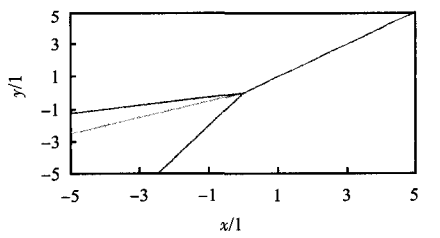

图 4　PReLU 函数

与 ReLU 函数相比,PReLU 函数中的负半轴斜率系数 α 可以学习而非固定,输出均值趋近于 0,而且 $x<0$ 时函数非硬饱和,因此,PReLU 函数的收敛速度更快,无神经元死亡的问题。

另外,其他激活函数如 RReLU,ELU 等亦能够提高收敛速度。

2　改进的 ReLU 激活函数 TLU

对 ReLU 函数进行了改进,将 ReLU 函数 $x<0$ 的部分使用 tanh 函数代替,构造出了一个新的激活函数 TLU,函数定义为

$$f(x)=\begin{cases}x, x\geqslant 0\\ \alpha\tanh(x), \text{其他}\end{cases} \tag{5}$$

其图像如图 5,其中,$x<0$ 部分图像根据斜率 α 变化。

从函数表达式和图像中可以看出,TLU 在右侧的线性

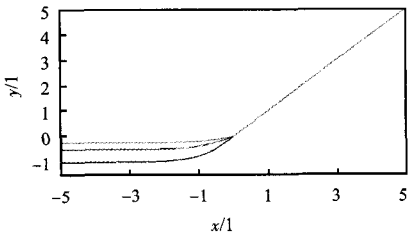

图5 TLU函数

部分具有函数ReLU和LReLU的优点，在$x \geqslant 0$时导数为常数，因此，在饱和区内的梯度永远不会为0，能够有效缓解梯度消失问题。

1)TLU函数与ReLU函数对比，左侧的非线性部分($x<0$部分)不仅能够使得均值更接近于0，避免均值偏移现象，而且由于其左侧部分不具备硬饱和的性质，TLU不会出现神经元死亡现象。

2)虽然LReLU函数在$x<0$部分也能取值从而使均值趋近于0，但LReLU函数左侧部分是线性的，对输入变化或噪声的鲁棒性较弱，而TLU函数左侧部分是非线性的具有软饱和性，鲁棒性更好，因此，可以预测TLU函数的性能必然强于LReLU函数。

3 实验与结果分析

采用字符集语言模型，在GRU型的深度网络结构进行实验。实验环境为Ubuntu15.04LTS，Torch7，LuaRocks以及使用NVIDIA推出的通用并行计算架构CUDA Toolkit的NVIDIA GPU。训练数据集是部分Linux Ubuntu源代码，约5 MB。

由于使用sigmoid激活函数进行实验时出现了结果不收敛的情况，所以实验结果仅使用了同为sigmoid系的tanh函数作为对照组，另外还增加了一个ELU激活函数作为对照组。使用不同的激活函数的GRU型深度神经网络的训练结果如图6所示，实验结果表明：在相同的训练时间下，误差率从低到高排序依次为TLU<ELU<LReLU<Tanh，使用TLU函数的作为激活函数可显著地减少训练误差。

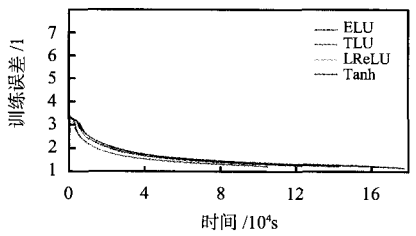

图6 实验结果

另一方面，在同等误差率下，按照训练时间从小到大排序依次为TLU<ELU<LReLU<Tanh，说明在同等误差率下，使用TLU函数作为激活函数的实验，其训练时间明显较另外3组实验少。

4 结束语

设计了一种新的激活函数TLU，并与一些sigmoid系和ReLU系的激活函数进行了比较，实验证明：TLU能显著地加快深度神经网络的训练速度并有效地降低训练误差。实验表明，TLU的系数α对训练时间和误差有一定的影响，下一步研究工作将对参数α进行优化，以进一步提高TLU函数的性能。

参考文献：

[1] Gers F. Long short-term memory in recurrent neural networks[D]. Hannover, Germany: Universität Hannover, 2001.

[2] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv: 2014, 1406.1078.

[3] 李宏伟，吴庆祥. 智能传感器中神经网络激活函数的实现方案[J]. 传感器与微系统，2014，33(1)：46-48.

[4] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines[C]// Proceedings of the 27th International Conference on Machine Learning(ICML), 2010: 807-814.

[5] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6(2): 107-116.

[6] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Neural Information Processing System, 2012: 1097-1105.

[7] Clevert D E, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus)[J]. arXiv preprint arXiv: 2015, 1511.07289.

[8] Gulcehre C, Moczulski M, Denil M, et al. Noisy activation functions[J]. arXiv preprint arXiv: 2016, 1603.00391.

[9] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[Z]. 2015: 1026-1034.

作者简介：

蒋昂波(1991-)，男，通讯作者，硕士研究生，研究方向为深度学习、神经网络。

王维维(1957-)，男，副教授，硕士生导师，主要从事电子设计自动化(EDA)、计算机体系结构、人工智能研究工作。

(上接第49页)

[6] 任志平，李 躭，党 博. 基于PIC单片机的找水系统设计[J]. 传感器与微系统，2016，35(10)：73-75.

[7] 徐建华，刘迪仁. 绕于芯棒上的电流环在多层环状媒质中的电磁响应[J]. 电子学报，1999，27(6)：9-12.

[8] Sakaji N M. Force and eddy currents in a solid conducting cylinder due to an eccentric circular current loop[J]. Journal of Physics D: Applied Physics, 2000, 33(18): 1-11.

[9] 美国国家仪器(NI)有限公司. USB-4431 数据采集卡[EB/OL]. http://sine.ni.com/nips/cds/view/p/lang/zhs/nid/206676

作者简介：

刘红兰(1970-)，女，高级工程师，主要从事浅海油气田采油工艺研究及应用推广工作，E-mail：liuhonglan.slyt@sinopec.com。