·人工智能与模式识别·　　文章编号：1000-3428(2022)07-0104-10　　文献标志码：A　　中图分类号：TP18

深度学习中的权重初始化方法研究

邢彤彤，孙仁诚，邵峰晶，隋　毅

（青岛大学 计算机科学技术学院，山东 青岛 266071）

摘　要：深度神经网络训练的实质是初始化权重不断调整的过程，整个训练过程存在耗费时间长、需要数据量大等问题。大量预训练网络由经过训练的权重数据组成，若能发现预训练网络权重分布规律，利用这些规律来初始化未训练网络，势必会减少网络训练时间。通过对AlexNet、ResNet18网络在ImageNet数据集上的预训练模型权重进行概率分布分析，发现该权重分布具备单侧幂律分布的特征，进而使用双对数拟合的方式进一步验证权重的单侧分布服从截断幂律分布的性质。基于该分布规律，结合防止过拟合的正则化思想提出一种标准化对称幂律分布（NSPL）的初始化方法，并基于AlexNet和ResNet32网络，与He初始化的正态分布、均匀分布两种方法在CIFAR10数据集上进行实验对比，结果表明，NSPL方法收敛速度优于正态分布、均匀分布两种初始化方法，且在ResNet32上取得了更高的精确度。

关键词：深度学习；卷积神经网络；预训练模型；权重初始化；对称幂律分布

开放科学（资源服务）标志码（OSID）：

中文引用格式：邢彤彤，孙仁诚，邵峰晶，等.深度学习中的权重初始化方法研究[J].计算机工程，2022，48(7)：104-113.

英文引用格式：XING T T，SUN R C，SHAO F J，et al.Research on weight initialization method in deep learning[J].Computer Engineering，2022，48(7)：104-113.

Research on Weight Initialization Method in Deep Learning

XING Tongtong，SUN Rencheng，SHAO Fengjing，SUI Yi

（School of Computer Science and Technology，Qingdao University，Qingdao，Shandong 266071，China）

【Abstract】The essence of deep neural network training is the constant adjustment of the initial weight，and the entire training process is time consuming and requires a large amount of data.Most pretraining networks are essentially trained weight data. If the weight distribution rules of pretraining networks are identified and untrained networks can be initialized using these rules，then the network training time can be reduced. In this study，the probability distribution analysis of the pre-training model weights of AlexNet and ResNet18 on the ImageNet dataset is performed；the result shows that the weight distribution exhibits the characteristics of a one-sided power law distribution. Subsequently，the double logarithm fitting method is used to verify that the one-sided distribution of weight obeys the truncated power law distribution.Combining the distribution law with the regularization idea to prevent overfitting，an initialization method fora Normalized Symmetric Power Law（NSPL）distribution is proposed. Subsequently，the normal and uniform distribution methods initialized by He on the AlexNet and ResNet32 networks are compared experimentally on the CIFAR10 dataset.The experimental results show that the convergence rate of the NSPL distribution initializing method is higher than those of the two abovementioned initializing methods，and that ResNet32 achieves higher accuracy.

【Key words】deep learning；Convolutional Neural Network（CNN）；pre-training model；weight initialization；symmetric power law distribution

DOI：10. 19678/j. issn. 1000-3428. 0062017

0　概述

从MCCULLOCH等[1]提出神经网络的初步概念以及神经元的数学模型开始，针对神经网络的研究得到迅速发展。特别是深度学习[2]在图像领域的优秀表现，使其在机器学习[3]中脱颖而出。其中，卷积网络[4]的概念也越来越受到人们的关注和重视，尤其是在图像分类的处理中，卷积神经网络的表现非常突出。尽管训练越来越深的网络存在一些困难，但是卷积神经网络还是取得了较好的成绩，并且

基金项目：国家自然科学基金青年科学基金项目（41706198）。

作者简介：邢彤彤（1997—），女，硕士研究生，主研方向为深度学习、大数据；孙仁诚（通信作者），教授、博士；邵峰晶，教授、博士、博士生导师；隋　毅，副教授、博士。

收稿日期：2021-07-08　　**修回日期：**2021-09-06　　**E-mail：**deniecex@163.com

还在不断的优化、突破。

深度学习的本质就是学习、优化权重的值，使其达到一个最优解的状态。通过文献[5]提出的卷积神经网络可视化方式可清楚地观察到卷积神经网络每一层的权值情况，这其中需要更新权重的层，包括卷积层、BN层和FC层等。在寻找最优解的过程中，权重的初始化就是得到最优解的重要前提。如果权重初始化不合适，则可能会导致模型反向传播[6]失效，陷入局部最优解，使得模型预测效果不理想，甚至使损失函数震荡，模型无法收敛，即使用不同的权重初始化方法，能够直接影响模型的训练速度和最终精确度。因此，一个优秀的权重初始化方法是模型提升收敛速度和最终精确度的重要前提。

在深度学习领域中，卷积神经网络的权重初始化可以采取多种方式，如高斯（正态）分布初始化[7]、均匀分布初始化[8]、截断高斯分布初始化[9]（该初始化方法与高斯分布初始化相似，但分布形式为截尾分布）以及主成分洗牌初始化[10]等方法。其中，目前较为流行的权值初始化方法，如Xavier初始化方法[11]和He初始化方法[12]是在正态分布和均匀分布的基础上进行了改进。Xavier初始化为了增加网络各层之间信息传播的流畅性，遵循了（正向传播）各层激活值方差和（反向传播）各层状态值的梯度方差在传播中保持一致的原则，通过均匀分布来进行权重初始化调整。He初始化在Xavier初始化的基础上稍加改变，遵循（正向传播）各层状态值方差和（反向传播）各层激活值的梯度方差在传播中保持一致的原则，在与ReLU激活函数[13]的共同作用下，可以得到较好的收敛效果。然而，使用这两种权重初始化方法的网络依然存在训练时间长、需要数据量大的问题。文献[14]在实验过程中发现预训练模型[15]的权重参数分布可能存在幂律分布的现象，经过其后期验证得出预训练权重存在局部幂律的性质。

本文从Pytorch中图像分类相关的预训练模型[16]入手，分析预训练模型的权重分布，提出一种标准化的对称幂律（Normalized Symmetric Power Law，NSPL）初始化方法。分析权重初始化面临的主要问题，研究预训练网络模型的权重分布，发现权重分布具备幂律分布的特征。在此基础上，基于标准化对称幂律分布，给出权重数据生成及初始化算法。

1　问题描述

1.1　权重初始化问题

权重有效初始化可以防止激活值在深度神经网络的正向传递过程中出现梯度爆炸或者梯度消失。模型经过权重初始化后，在训练、更新权重时主要会出现以下2种情况：

1）如果初始权重太小，导致神经元的输入过小，随着层数的不断增加，会出现信号消失的问题，也会导致sigmoid激活函数[17]中强调的丢失非线性的能力，因为在0附近sigmoid函数是近似线性的。

2）如果初始权重太大，会导致输入状态也较大，对sigmoid激活函数来讲，激活函数的值会变得饱和，从而出现梯度消失的问题。

无论上述哪一种情况发生，损失梯度要么太大或要么太小，更新信息都无法有效地向后传递，网络则需要很长时间才能收敛。研究人员研究了各种初始化方法来避免这些问题，如：通过保持一层网络的输入和输出方差不变来防止梯度消失的Xavier初始化方法；He初始化方法通过加重权重方差的方式弥补ReLU激活函数[18]1/2为零的状态。

目前针对权重初始化方法的思路更多偏向于正态分布和均匀分布，但还不能更好地以合适的数据对深度学习网络进行初始化。若使模型的初始权重分布与训练后模型权重的分布接近，将有助于模型获得最优解，减少模型的训练时间。因此，寻找一个更合适的数学分布规律来进行权重初始化，是本文探讨并验证的核心问题。

1.2　权重初始化方法

网络模型的训练实质就是更新权值并找到最优权值的过程。预训练模型的权值就是网络训练最终找到的最优权值，若可以从预训练模型的权值中总结出规律，研究并制定一种权重初始化的方法，有助于提升网络模型的训练速度和最终精确度。

针对权重初始化目前存在的问题，本文提出一种有效的权重初始化方法，具体解决思路如下：1）从预训练模型的权值入手，查看并分析预训练模型的权值分布规律；2）通过分析预训练模型的权值分布特征，发现权重分布具有幂律分布特征，进一步进行幂律分布拟合的检验实验，考虑制定一种以幂律分布为基础的权重初始化方法；3）优化数据分布结构，制定标准化的对称幂律分布数据，即本文提出的NSPL初始化方法。

本文从预训练模型的权重入手，查看并分析预训练模型的权重分布规律，探究幂律分布在权重初始化中的作用。对比实验结果表明，本文提出的方法有助于减少网络权重的训练时间，具有提升网络最终精确度的能力。

1.3　预训练模型的权重分布分析

本节使用的是Pytorch框架下torchvision中的预训练模型，它是基于ImageNet数据集上训练出来的，通过查看预训练模型的权值，对预训练模型权值做相关统计分布分析。依据幂律分布的判断性质，在双对数坐标下，幂律分布表现为一条斜率幂指数为负数的直线，这一线性关系是判断给定的实例中随机变量是否满足幂律的依据。本文对AlexNet和ResNet18预训练模型的所有卷积层权重进行双对数线性拟合，并计算其拟合优度R^2。

首先针对AlexNet[19]网络的卷积层权值分布进行处理。依次读取AlexNet预训练模型的权重参数，并使用概率分布来可视化权值的分布情况。该网络的五层卷积层权值数据的概率分布情况如图1所示，其预训练模型权重的双对数拟合图如图2所示。根据AlexNet

预训练模型的权值分布情况，可以通过其高峰、长尾的特点，进一步对更深层的ResNet18[20]预训练模型进行相同的实验。其中，ResNet18共有17层（加上输入层）卷积层，其权值概率密度分布情况如图3所示，其预训练模型权重的双对数拟合图如图4所示。

从图1和图3可以看出，这两个网络的预训练模型权值数据皆具有高峰、长尾的特点。在各种数学分布中，同样具有该特点的是幂律分布，推断预训练模型的权重分布单侧倾向幂律分布。从图2和图4中的双对数线性拟合结果可以看出，所有层权重线性拟合优度R^2值都是在0.8左右，可以得出预训练网络模型的权重分布并不完全为幂律分布，属于指数截断的幂律分布[21]。从数据上来看，实际分布中权值接近于0的数据少于幂律分布，但根据对深度网络模型正则化[22]研究结果，在损失函数中加入L1或L2正则化项[23]，将使模型中更多的权值为0或者接近于0，且模型的泛化能力更强。基于此，本文以幂律分布来初始化网络，而没有采用指数截断的幂律分布。本文制定一种标准化的对称幂律分布的权重初始化方法，用来确定幂律分布在权值中的作用。

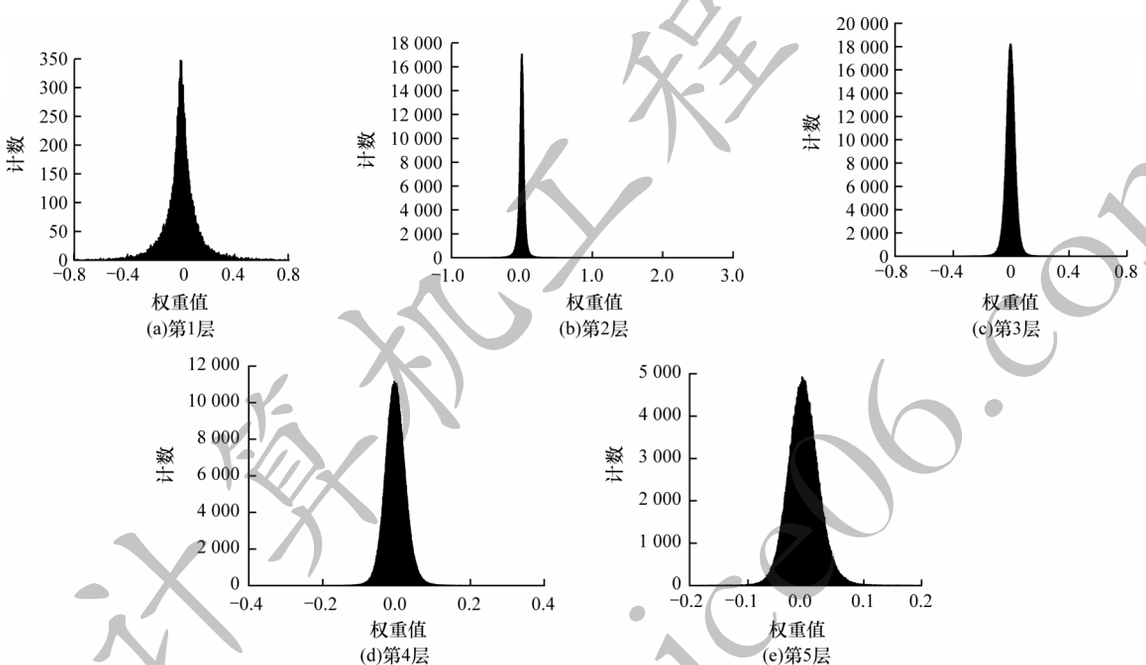

图1 AlexNet预训练模型权重数据概率分布

Fig.1 Probability distribution of weight data of AlexNet pre-training model

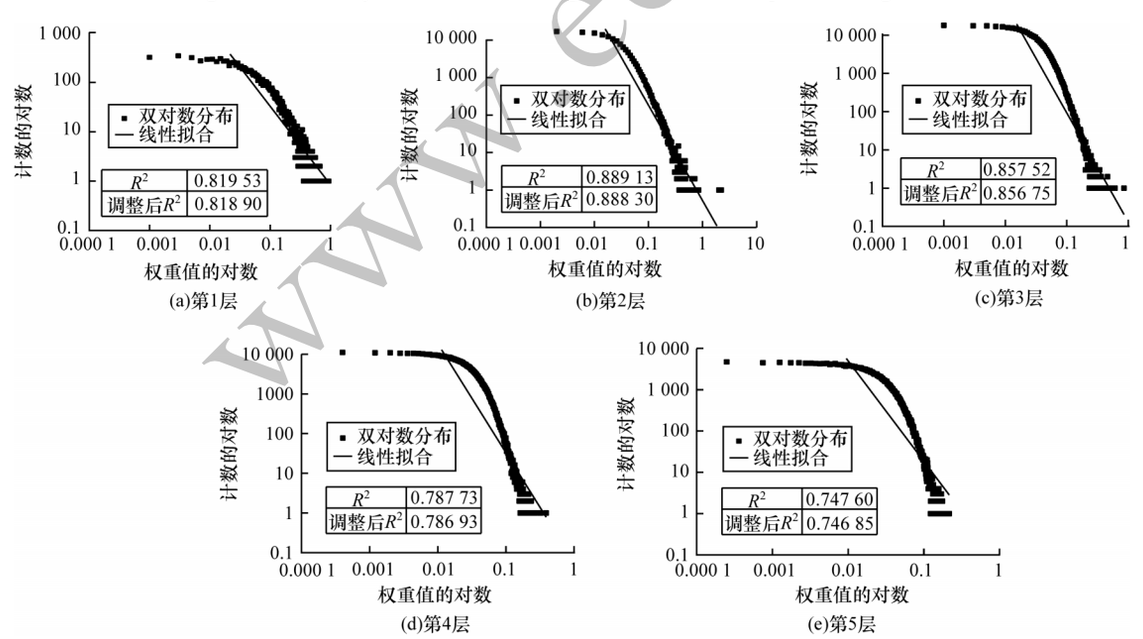

图2 AlexNet预训练模型权重的双对数拟合图

Fig.2 Double log-fitting diagram of AlexNet pre-training model weight

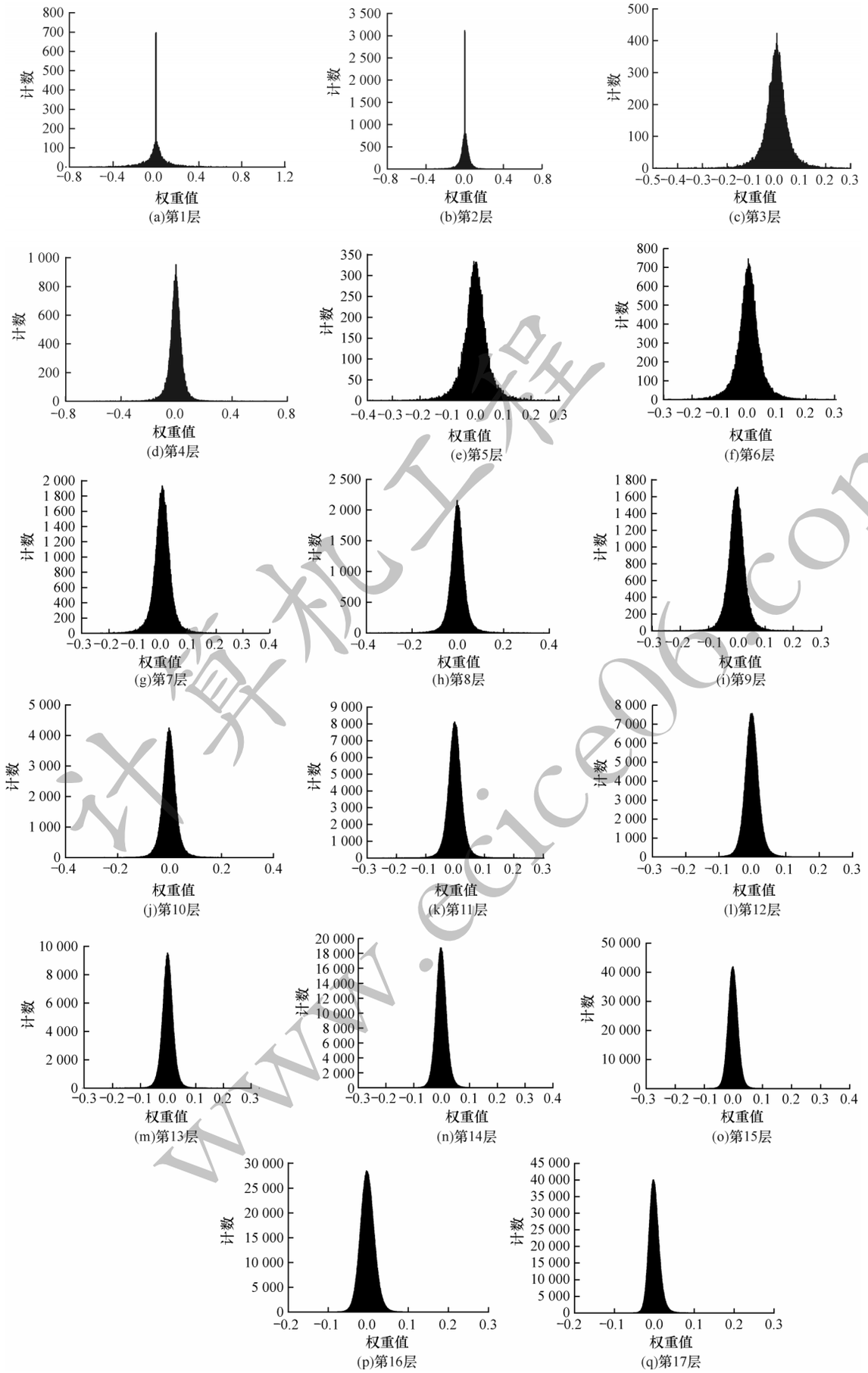

图3　ResNet18预训练模型权重数据概率分布

Fig.3　Probability distribution of weight data of ResNet18 pre-training model

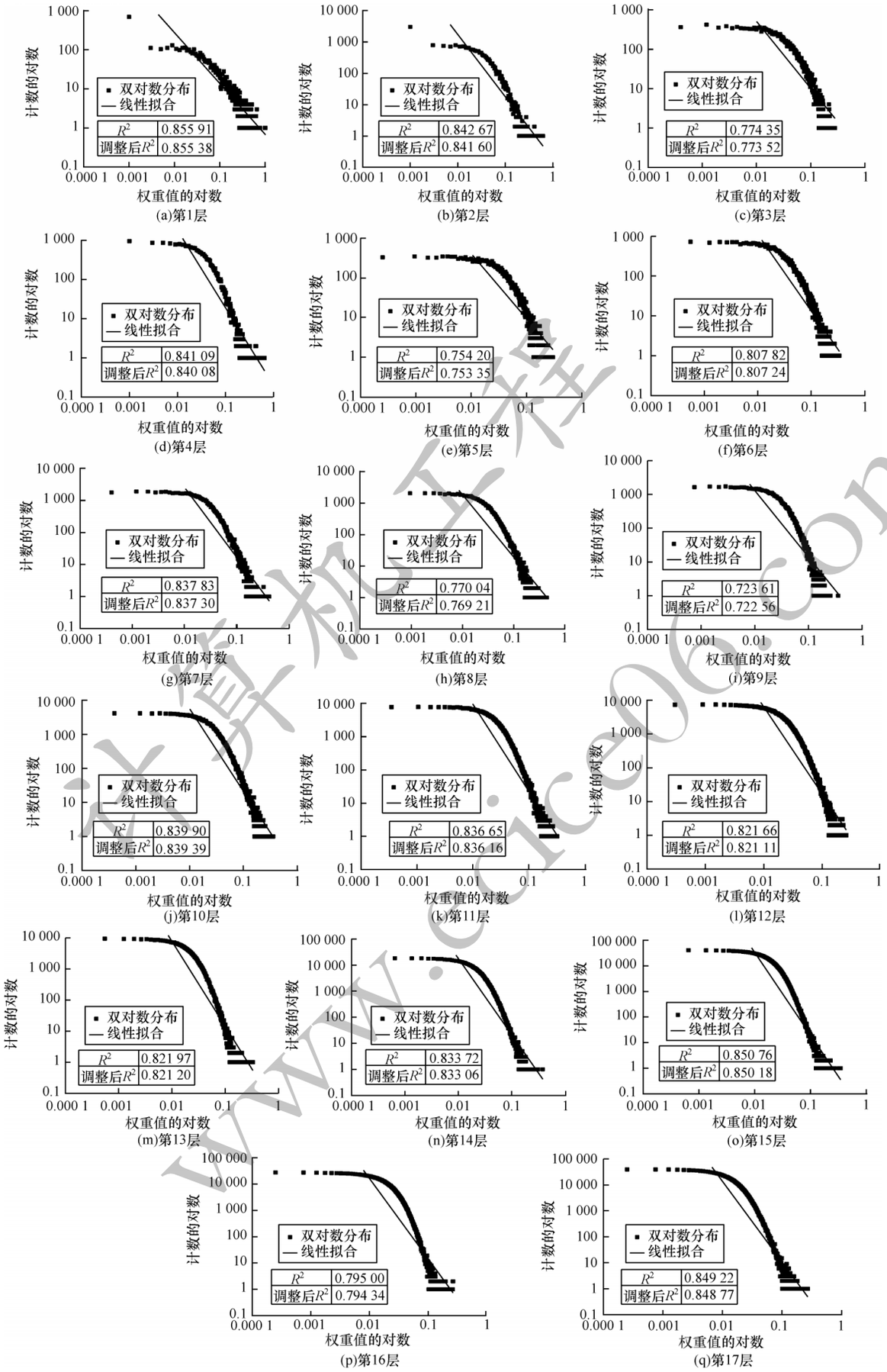

图4 ResNet18预训练模型权重的双对数拟合图

Fig.4 Doublelog-fitting diagram of ResNet18 pre-training model weight

2　标准化对称幂律分布

2.1　函数形式的数学推导

归一化的对称幂律函数推导过程如下：

1）幂律分布的公式为：

$$f(x)=cx^{-\alpha},c>0,\alpha>0 \tag{1}$$

2）标准化过程，令：

$$\int_{x_{\min}}^{\infty} f(x)=1 \tag{2}$$

通过计算得：

$$c=\frac{\alpha-1}{x_{\min}^{1-\alpha}} \tag{3}$$

将c代入式（1）得：

$$p(x)=\frac{\alpha-1}{x_{\min}^{1-\alpha}}x^{-\alpha},x\geqslant x_{\min},\alpha>1 \tag{4}$$

式（4）即为单侧标准化的幂律分布函数公式。

标准化对称幂律函数为：

$$p(x)=\frac{\alpha-1}{x_{\min}^{1-\alpha}}\left(x\times\operatorname{sign}(x)\right)^{-\alpha},x\geqslant x_{\min},\alpha>1 \tag{5}$$

2.2　权值数据的生成算法

本节算法致力于生成指定数量的标准化对称幂律数据，用来初始化不同网络模型，具体步骤如下：

步骤1　根据幂律分布的公式，推导出标准化的幂律函数（见2.1节）。

步骤2　将标准化的幂律函数做对称，得到标准化的对称幂律函数。

步骤3　分别计算网络模型中各个卷积层的参数数量。

步骤4　运用本文算法得到对应数量的参数值，分别对网络的卷积层权值重新初始化。生成的权值应当符合分布要求，并且无大量、连续的相同数据。

1）对称幂律函数的生成算法

依据算法设计，算法1可以得到标准化的对称幂律分布函数。

算法1　标准化的对称幂律函数

输入　N,α,ex；//N为图像描点数，α为归一化参数，ex//为跨度值

输出　X,Y；//XY为一组标准化的对称幂律函数集合

1.N，Half N=N/2；

2.α，C = 1 − α；

3.for x = 1，2，…，half N + 1

4.$y=\frac{-C}{x_{\min}^{C}}x^{C-1}$；

5.end for

6.$y.sum=\sum_{i=1}^{500}(y^{i})$；

7.y.sum=1；//y值归一化；

8.$x_1=\frac{-x}{ex},x_2=\frac{x}{ex}$；

9. X=（x_1，x_2）；

10.$y_1=\frac{-y}{2},y_2=\frac{y}{2}$；

11. Y=（y_1，y_2）；

12. $Y.sum=\sum_{i=1}^{500}(Y^{i})$；

13.Y.sum = 1；//Y值归一化

14. return X，Y；

2）对称幂律数据生成

在算法1建立了标准化的对称幂律分布函数后，根据计算出的网络模型每一层的权重参数量，使用算法2来生成对称幂律数据。

算法2　对称幂律数据生成算法

输入　X,Y,num,m//XY为一组标准化的对称幂律函//数集；num为网络层需要的参数量；m为随机数参数

输出　powerlaw//NSPL初始化规则的网络参数

1.$a_i=Y_i$num，(i = 1，2，…，X.size)；

2. for i = 0 to X.size do：//X.size为X集合大小

3.A；//A为新的空数列

4.count = 0，needcount=a_i；//count计数

5.while count<needcount do：

6.randnum；//randnum为（−m，m）的随机数；

7.if A == empty or randnum！ =A.last：//A.last为A数列//最后一个元素

8.randnum is add to list A；

9.for j = 0：a_i do：

10.A =A.values +x_i；//A.values为A中每一个元素

11.A is add to powerlaw；

12.return powerlaw

3　实验结果与分析

为验证本文提出的NSPL初始化方法有助于缩短网络训练时间，提高网络的最终精确度，设置以下的对比实验：运用cifar10数据集分别在AlexNet网络和ResNet-32网络上进行训练，而在训练过程中每一个网络都将使用3种权重初始化方法进行初始化，分别为He的均匀分布初始化、He的正态分布初始化[16]和NSPL初始化。

通过实验分别对比同一个网络下不同初始化方法的初始精确度和最终模型精确度的差异，最终得出本文提出的NSPL初始化可以有效提高模型的训练速度和最终精确度。

本文两组实验的流程设计如下：

1）获取数据集，设置网络模型；

2）计算并记录每一层网络模型参数量；

3）利用算法1制作出标准化的对称幂律分布函数；

4)利用算法2生成与网络模型参数量对应大小的对称幂律数据;

5)使用3种不同的权重初始化方法对网络模型的参数进行初始化;

6)使用训练集进行训练,学习权重参数;

7)每一轮训练集结束后,使用验证集进行准确率验证,并记录该准确率。

3.1 卷积层参数量计算

卷积层权重参数量的计算公式为:

$$weights=kernel\_size^2\times in\_channels\times out\_channels \tag{6}$$

其中:in_channels表示输入的通道数;out_channels表示输出的通道数;kernel_size表示卷积核的大小。

1) AlexNet网络各层权重数量

结合式(6)计算AlexNet网络所有卷积层的权重数量,如表1所示。

表1 AlexNet网络各层权重数量

Table 1 Weight quantity of each layer of AlexNet network

卷积层	权重数量计算式	权重数量
features.0.weight	96×3×7×7	14 112
features.3.weight	256×96×5×5	614 400
features.6.weight	384×256×3×3	884 736
features.8.weight	384×384×3×3	1 327 104
features.10.weight	256×384×3×3	884 736

2) ResNet32网络各层权重数量

ResNet32网络是以block块为基本单位组成的网络结构,因此在此处以不同的block来对不同的卷积层的情况进行描述。结合式(6)计算该网络卷积层种类以及对应的权重参数量,如表2所示。

表2 ResNet-32网络各层权重数量

Table 2 Weight quantity of each layer of ResNet-32 network

卷积层	权重数量计算式	权重数量
conv1_1	64×3×7×7	9 408
conv2_6	64×64×3×3	36 864
conv3_1	128×64×3×3	73 728
conv3_7	128×128×3×3	147 456
conv4_1	256×128×3×3	294 912
conv4_11	256×256×3×3	589 824
conv5_1	512×256×3×3	1 179 648
conv5_5	512×512×3×3	2 359 296

3.2 网络初始权重分布情况

下文所有权重数据与ResNet32网络相似,此处仅以AelxNet为示例。

1) NSPL初始化数据。使用本文提出的算法结合AlexNet的五层卷积层所需要的权重参数量,生成NSPL初始化数据。本文算法生成的权重初始化数据分布如图5所示。从图5可以看出,该数据充分展现了幂律分布的高峰、长尾现象。因为是标准化的对称幂律分布,所以高峰和长尾特征比较明显。

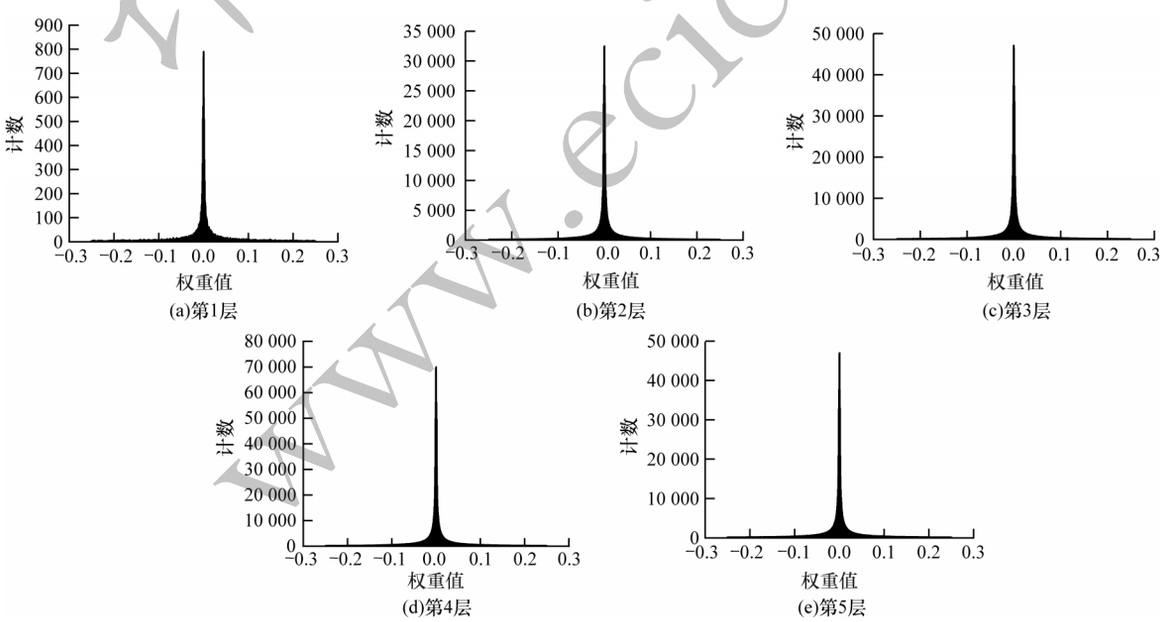

图5 对称幂律初始化数据分布

Fig.5 Distribution of symmetric power law initialization data

2)He方法的正态分布初始化数据。使用He正态分布初始化方法对网络权重进行初始化,读取网络初始权值,将其数据分布可视化,如图6所示。该方法的权值数据分布就是非常典型的正态分布钟形图像,依据该方法对方差的计算和控制可以看出每一层数据分布情况的不同。

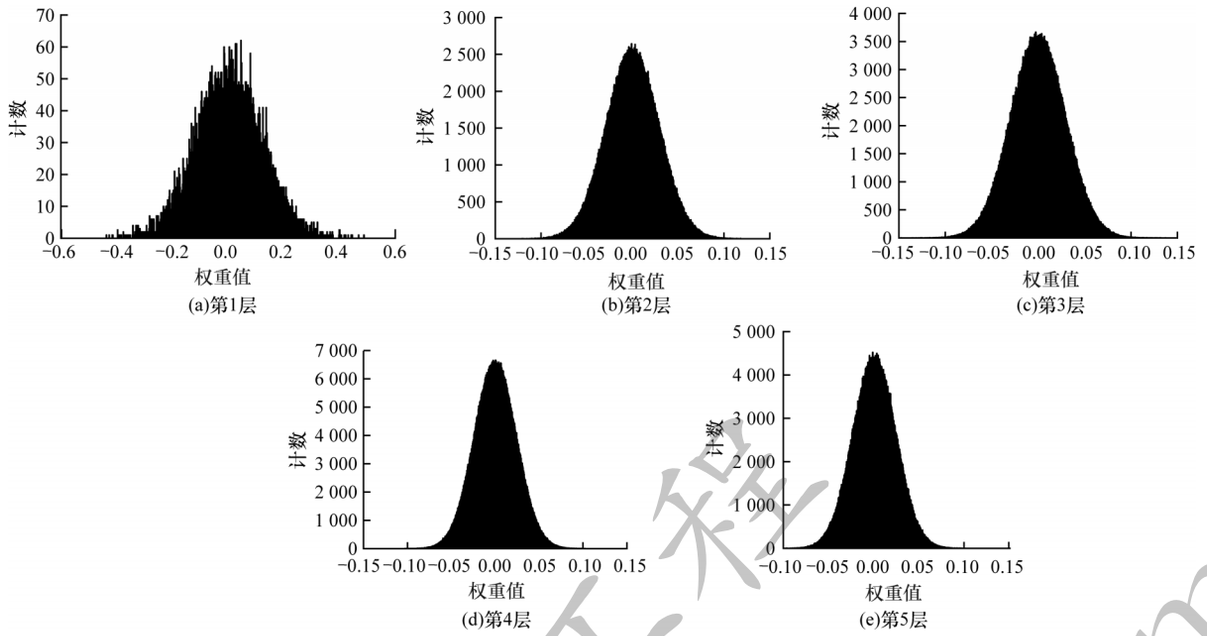

图6　He方法的正态分布初始化数据分布

Fig.6　Distribution of normal distribution initialization data of He method

3)He的均匀分布初始化数据。使用He均匀分布初始化方法对网络权重进行初始化,读取网络初始权值,也就是该初始化方法生成的数据,该权重初始化方法的数据分布如图7所示。该初始化方法是Pytorch1.7中默认的初始化方法,当网络不指定初始化方法时,会调用该方法对卷积层进行初始化。

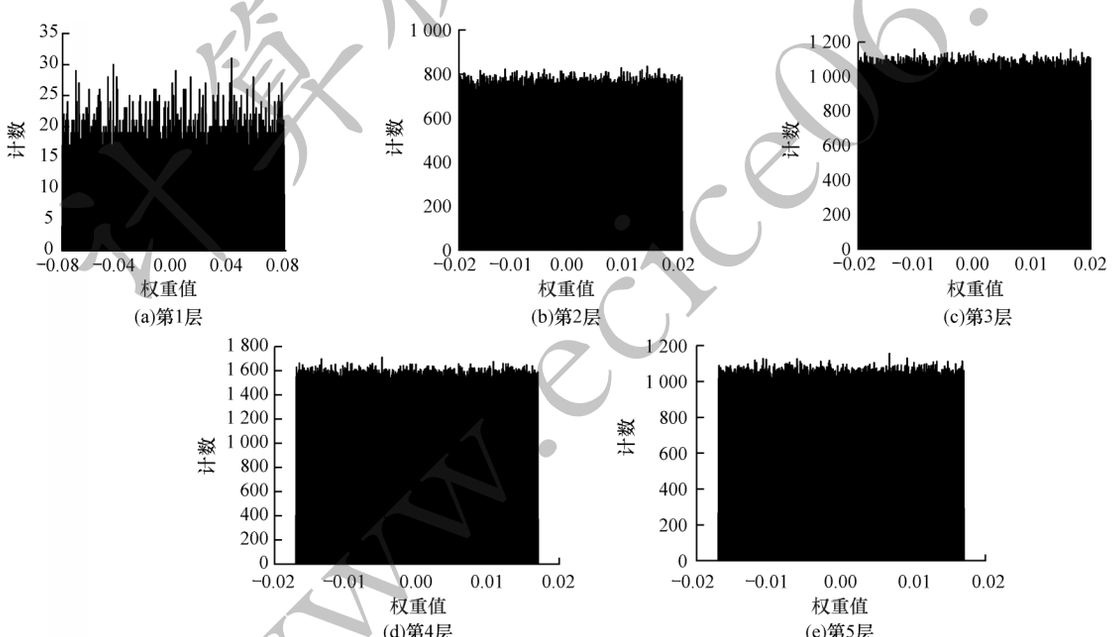

图7　He方法的均匀分布初始化数据分布

Fig.7　Distribution of uniformly distributed initialization data for He method

3.3　对比实验

对比实验过程如下:

1) 实验设计

本文实验使用cifar-10数据集,在AlexNet网络和ResNet32网络上进行实验,将NSPL初始化的实验结果与He的正态分布初始化、均匀分布初始化方法的实验结果进行对比分析。

cifar10数据集是一个更接近现实物品的RGB彩色图像,包含10个类别,每个类别有6 000个图像,分别为50 000张训练图片和10 000张测试图片。本文实验在训练集上进行模型训练,使用测试集进行测试,以对比测试集的准确度。

本文实验是对比使用不同权重初始化的网络训练首轮次训练后的测试集精确度及后续网络模型的收敛速度。通过对比同一训练轮次下的不同初始化方法达到的精确度,得出其中一个初始化方法

更有助于提升网络训练速度和最终模型准确率的结论。

2）实验过程

针对AlexNet网络和ResNet32网络，分别使用上文中提到的3种方法进行权重初始化。网络每一轮次训练结束都用验证集测试当前网络的准确度并进行记录，将3种精确度对应的所有轮次的验证集精确度进行对比分析。

（1）在AlexNet网络实验过程中，使用的超参数设置如下：随机梯度下降法（Stochastic Gradient Descent，SGD）优化器，动量momentum=0.9，批尺寸batch_size=64，学习率*lr*=0.015，测试尺寸test_batch=1 000，训练轮次epochs= 30，损失函数使用CrossEntropyLoss。

图8所示为3种不同权重初始化方法在AlexNet网络上各个轮次的训练精确度。

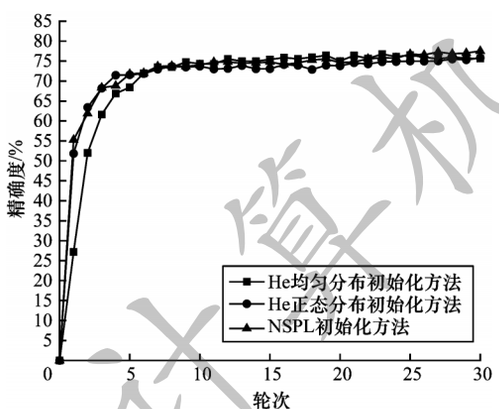

图8 AlexNet训练精确度对比

Fig.8 Comparison of AlexNet training accuracy

在AlexNet网络的对比实验中，通过图8可以看出NSPL初始化方法和He正态分布初始化的初始轮次精确度优于He均匀分布初始化，NSPL初始化方法相较于He的均匀分布和正态分布初始化方法的最终精确度也有微弱的提升。本文实验进一步使用了具有更高复杂度的ResNet32网络模型来验证NSPL初始化的使用效果。

（2）在ResNet32网络实验过程中，使用的超参数设置如下：SGD优化器，动量momentum=0.9，批尺寸batch_size=128，学习率为*lr*=0.01，测试尺寸test_epochs=100，训练轮次epochs = 30，损失函数使用CrossEntropyLoss。

图9所示为3种不同权重初始化方法在ResNet32网络上各个轮次的训练精确度对比，通过图9可以看出，在模型精确度提升的过程中，NSPL初始化有助于优化网络的训练过程，加快收敛速度。

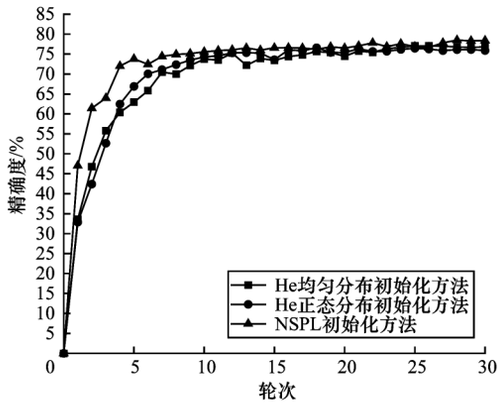

图9 ResNet32训练精确度对比

Fig.9 Comparison of ResNet32 training accuracy

3.4 对比实验分析

通过图8实验结果对比可以发现，He的正态分布初始化方法和本文提出的NSPL始化方法在初始轮次中有较高的准确度。在最终模型趋于稳定时，NSPL初始化的精确度比He的正态分布初始化方法提高3%。总地来说，NSPL初始化在AlexNet网络上具有优化网络模型训练过程的优点。

通过图9实验结果对比可以发现，在更为复杂的ResNet32网络中，NSPL初始化方法在首轮次中的精确度比He初始化方法的精确度提高60%，并且模型收敛的速度更快，其最终精确度比He初始化方法提高8%。在更深层的网络中，NSPL初始化方法具有更优秀的表现。

通过上述两组对比实验可以发现，NSPL初始化方法有助于提升网络训练的速度和最终准确度，说明幂律分布也可以作为一种权重初始化的模型方法。

4 结束语

本文通过理论推导和实验验证，提出一种提升网络模型训练速度和精确度的权重初始化方法——对称幂律（NSPL）初始化方法。同时，设置2种网络结构，在3种不同权重初始化下进行对比实验，使用cifar10数据集分别训练，对比每一轮次的模型训练精确度。实验结果表明，本文NSPL初始化方法能够优化网络训练过程，加快收敛速度。本文采用的是标准化后的对称幂律数据，并没有深入研究截断幂律分布拟合的情况，下一步将统计并分析大量预训练模型的权重参数分布情况，结合不同网络模型的层数及不同数据集等影响权重初始化的因素，制定出更有针对性的基于幂律分布的初始化方法。

参考文献

[1] MCCULLOCH W S, PITTS W. A logical calculus of the ideas immanent in nervous activity [J]. Bulletin of Mathematical Biology, 1990, 52(1/2): 99-115.

[2] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[3] ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain [J]. Psychological Review, 1958, 65(6): 386-408.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] 司念文，张文林，屈丹，等. 卷积神经网络表征可视化研究综述[J/OL]. 自动化学报：1-31[2021-02-12]. https://doi:10. 16383/j. aas. c200554.
SI N W, ZHANG W L, QU D, et al. A review on representation visualization of convolutional neural networks[J]. ActaAutomatica Sinica: 1-31[2021-02-12]. https://doi:10. 16383/j. aas. c200554. (in Chinese)

[6] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[7] 刘晴. 一种改进的深度卷积神经网络及其权值初始化方法研究[D]. 保定：河北大学，2018.
LIU Q. An improved deep convolutional neural network and its weight initialization[D]. Baoding: Hebei University, 2018. (in Chinese)

[8] 沈成恺. 卷积神经网络权值初始化方法研究[D]. 北京：北京工业大学，2017.
SHEN C K. Research on initialization method of convolutional neural networks [D]. Beijing: Beijing University of Technology, 2017. (in Chinese)

[9] BURKARDT J. The truncated normal distribution[EB/OL]. [2021-06-01]. https://www. doc88. com/p-1176985733398. html.

[10] 李玉鑑，沈成恺，杨红丽，等. 初始化卷积神经网络的主成分洗牌方法[J]. 北京工业大学学报，2017，43(1)：22-27.
LI Y J, SHEN C K, YANG H L, et al. PCA shuffling initialization of convolutional neural networks[J]. Journal of Beijing University of Technology, 2017, 43(1): 22-27. (in Chinese)

[11] SHEN H. Towards a mathematical understanding of the difficulty in learning with feedforward neural networks [EB/OL]. [2021-06-01]. https://arxiv. org/abs/1611. 05827.

[12] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2015: 1026-1034.

[13] 张焕，张庆，于纪言. 激活函数的发展综述及其性质分析[J]. 西华大学学报(自然科学版)，2021，40(4)：1-10.
ZHANG H, ZHANG Q, YU J Y. A review of the development and property analysis of activation function [J]. Journal of XihuaUniversity(Natural Science Edition), 2021, 40(4): 1-10. (in Chinese)

[14] 李杰. 卷积神经网络的权重初始化研究及应用[D]. 青岛：青岛大学，2020.
LI J. Research and application of weight initialization of convolutional neural networks [D]. Qingdao: Qingdao University, 2020. (in Chinese).

[15] HAN X, ZHANG Z Y, DING N, et al. Pre-trained models: past, present and future[J]. AI Open, 2021, 2: 225-250.

[16] KETKAR N S. Introduction to PyTorch[M]. Germany, Germany: Springer, 2017.

[17] HAN J, MORAGA C. The influence of the sigmoid function parameters on the speed of backpropagation learning[C]//Proceedings of IEEE International Workshop on Artificial Neural Networks. Washington D. C., USA: IEEE Press, 1995: 195-201.

[18] DAHL G E, SAINATH T N, HINTON G E. Improving deep neural networks for LVCSR using rectified linear units and dropout[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2013: 8609-8613.

[19] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[20] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.

[21] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[22] KANG G L, DONG X Y, ZHENG L, et al. PatchShuffle regularization[EB/OL]. [2021-06-01]. https://arxiv. org/abs/1707. 07103.

[23] MCMAHAN H B, HOLT G, SCULLEY D, et al. Ad click prediction: a view from the trenches[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2013: 1222-1230.

编辑　索书志